ENGR 40M course notes

September 2, 2020

Revision Notes:

1. Spring 2016: First revision of this reader created by the E40M teaching team of that quarter. Authors: Mark Horowitz, Steven Bell, Kristen Powell (ch. 3), Paul Lavengco(ch. 4), Christopher Ling (ch. 5), Theo Diamandis (ch. 6), Ya-Ting Wang (ch. 7), Miles Bennett (ch. 8).

2. Fall 2016: Minor revisions made to the impedance chapter to accommodate changes to the syllabus in that area.

3. Spring 2017: Practice problems and solutions to these problems can now be found at the end of each chapter. Minor revisions made.

4. Spring 2018: Revised sections 3.1, 3.2 (Nodal Analysis), 8.1, 8.2 (Op-Amps) and chapter 6 (Capacitors). Contributors: Chuan-Zheng Lee, Walker Ramirez.

5. Fall 2018: Major revision to the reader to add more examples and to align with current class. Currently finished with Chapter 1-7, Mark Horowitz

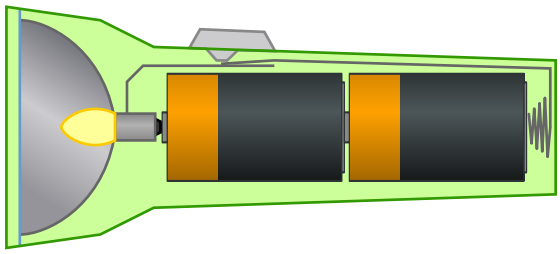
Contents

Chapter 1

What Makes Things Electrical?

Understanding Charge, Voltage, Current, and Power

In lecture we began by talking about a solar charger: a circuit that converts sun light to electrical energy. You will build one in your first lab. Since you probably haven't worked with a solar charger before, we will start our discussion of electrical systems with something more familiar: a flashlight. While you probably know how to use such a device (you flip a switch and the light comes on), and that it contains a number of different components, like batteries and a bulb, did you ever think of how it actually works? When you flip the switch the battery provides energy to the light bulb which causes it to light up, but how? In Section 1.2, we're going to take a brief look at each piece of the flashlight, and the following sections will then introduce you to the tools and terminology that are used to analyze electrical *circuits*. One of the keys to "making" is to understand how electrical circuits work, since they underlie most of the magical devices you use everyday. This chapter will introduce the concepts of charge, voltage and current, and show how they can be used to understand how power flows in an electrical circuit. Power flow is very important to flashlights, solar chargers, and all useful electrical circuits.



The figure above shows a cut away picture of an flashlight that uses 2 D-cell batteries, and even shows the metal "wires" that are used to carry the energy to the light bulb. The reason energy can flow is because of moving **charge** in the wires. But what is charge? In some ways electrical charge is the thing that makes electrical circuits electrical. Once we understand charge, we will look more closely at each of the electrical parts of a flashlight, and then look at how we represent a functional abstraction of each element in a circuit diagram (often called a schematic). Finally we will describe how to reason about a circuit's operation by using the concepts of voltage and current.

1.1 Electrical Charge

Charge, like mass, is a property of fundamental particles. You probably have heard/read at some point that an electron has a negative charge, a proton has a positive charge, and a neutron has no charge. Most of the time you don't need to worry about charge, since almost all large objects are charge neutral: atoms have the same number of electrons as protons, so stuff made up of atoms (nearly everything) is also charge neutral. So while you have heard about charge, it probably doesn't loom large in your thinking. Yet charge, especially moving charge, is fundamentally what makes electrical circuits electrical, so we will need to briefly look a little more closely at the properties of charge.

If you manage to have a number of charged particles in a space, say a number of electrons in a vacuum, there is a strong force that acts on them. This is the electrostatic force, one of the fundamental forces in nature. Its action is somewhat like gravity, except it is much, much stronger. Whereas gravity is related to mass, the electrostatic force is related to charge. Like gravity, the force between two charges is proportional to the magnitudes of the charges and decreases with distance, according to the equation:

$$F = k\frac{q_1 q_2}{r^2}$$

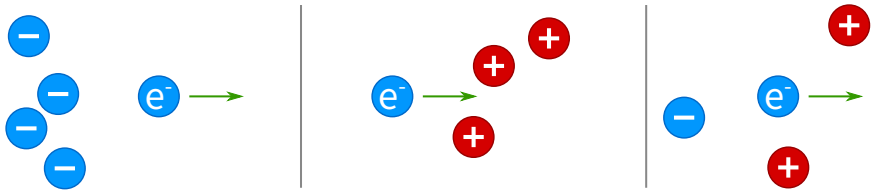
where F is the force between the two particles, q_1 is the charge of the first particle, q_2 is the charge of the second particle, and r is the distance between them. However, unlike mass for gravity, charges can be both positive and negative. So q can be either positive or negative. Notice that this means that the force can be either positive or negative. If the charge on both particles is the same, the force will be positive which will try to push the particles apart: like charges repel. However if the charges are of the opposite sign, the force will be negative so the force will try to pull the particles together: opposite charges attract. Two positive charges (or two negative charges) experience a force pushing them apart, while a negative and a positive charge experience a force pulling them together. The strong electrostatic force is the reason almost everything is charge neutral. Negative charged particles don't like to be together, and want to be with the opposite charge.

In the SI system, charge is measured in Coulombs, abbreviated "C". Most often, we'll be looking at electrons, which have a charge of -1.6×10^{-19} C, and protons, which have a charge of $+1.6 \times 10^{-19}$ C.

Suppose we observe an electron in space, accelerating to the right due to an electrostatic force.

Question: What configuration of charges could be causing this force?

One possibility is that there is a collection of negative charges on the left, pushing the electron away. Alternatively, there could be positive charges on the right, pulling the electron. Or there could be some combination of positive and negative charges, such that the sum of the attractive and repulsive forces is to the right.



Now image that we have a charge neutral object, which has the same number of + and - charges, and I start moving + charges to the top of the object, and don't let them flow back down, and move - charges to the bottom of the object, and again I don't let them flow back up. Notice that the object is still charge neutral, since I haven't changed the net charge of the device. Yet, as I increase the positive charge at the top, I will have to work against a larger electrostatic force to move the next positive charge to the top (since the charge in the top is larger), and the next negative charge to the bottom. It is like pumping water uphill, where the height increases as I add more charge.

Like the situation when you have pumped water up a hill, if someone provides an external path which allows the charge at the top of my object to flow to meet up with the negative charge at the bottom of the object, the charge will flow along this path. You can think of the charge flowing downhill, or the positive charges escaping from other positive charges to meet up with negative charges instead. It is energy I provided by separating the charge (pumping the water uphill) that charges "use" to flow through the external path. The force between charges (which can be represented by an electric field, but we won't go into that) and the ability of charges to move make electric circuits possible. To better understand how they work, let's look at a flashlight.

1.2 Inside a Flashlight

When you have a chance you should look inside of one of your flashlights. Most flashlights are very simple and only contain four different electrical *devices*: one or more batteries, wire, a switch, and a light emitting device. The latter used to be incandescent light bulbs, but now are often light-emitting diodes.

1.2.1 Battery

A battery is a chemical charge pump, using a pair of reduction-oxidation chemical reactions, one which produces negative charges (electrons) on a metal electrode, and a second that creates

positively-charged ions on the other electrode. Thus a battery uses chemical energy to separate charge in a charge neutral device. The chemical reactions generate a certain amount of energy per charge, and this energy is used to do the work needed to move the charge against the electrostatic force field (remember in physics moving something against a force does work, which means it takes energy). As you move charge, the force increases until you get to the point where the chemical reaction doesn't have enough energy to move any more charge (and increase the force). At this point the reaction stops. However if charge can flow from the battery (from both the + and - terminals) the chemical reaction will start again creating more charge. It is important to remember that the net charge in the battery (sum of + and - charges) is always zero, so the battery and all the devices we will discuss are charge neutral.

The energy per charge a battery can create is determined by the chemical reactions. This is measured in volts - we'll see why in the next section. Alkaline batteries are about 1.5 V, NiCd and NiMH are about 1.2 V, and Lithium polymer batteries are about 3.7 V.

Question: If all alkaline batteries have the same voltage, what is the difference between the different sizes of batteries?

Remember that every electron that flows out and back in to the battery is the result of a one-way chemical reaction. As the reagents get used up, the voltage drops until the battery is no longer useful. **Hence, since larger batteries physically hold more of the reagents, they can supply a larger number of electrons (higher current, and thus higher power) for a longer period of time (higher total energy).**

9-volt batteries are really just a package of six small cells wired together such that their total voltage is the sum of their individual voltages.

1.2.2 Wire

Materials can be electrically classified into one of three groups: *conductors*, through which charge can easily flow; *insulators*, which strongly resist charge motion; and *semiconductors*, which are in between. All three have their place in electronics. Wires, not surprisingly, are made of conducting materials. Insulators are used to keep the charge from going where it shouldn't, by separating things that would otherwise conduct. So most wires have a conducting core, and a plastic insulator coating. While the wire coating is the most obvious insulator example, there are many others: the fiberglass core of a printed circuit board and the stacks of ceramic disks which are used to suspend high-voltage power lines are also insulators. Finally, it turns out that semiconducting materials are extremely useful for building more complex electrical components. Small electrical changes can make a semiconductor go from insulator to conductor and back, which makes it possible to create devices that behave like switches or one-way valves for charge. We'll revisit these soon.

Now we're going to take a closer look at conductors, and examine what happens when we add charge to a conducting wire. Most metals are good conductors, and like all material consist of an array (a lattice) of atoms. Because of lots of physics that you may or may not want to know, some of the electrons of the metal atoms are not tightly bound to the nucleus and can wander around the metal. Remember, like the battery, the metal is still charge neutral (equal + and - charges), but the - charges are free to move.

Suppose we add some additional electrons to our conductor. Because like charges repel each other, this charge will push at the other charge in the wire, which will in turn push on other charge until some charge at the other end of the wire will flow out (if there is a place for the charge to

flow). It is similar to pushing water into a filled pipe. When you add water at one end, water flows out the other end almost instantly (but it is not the same water that you put in).

Conversely, if we add positive charge to one end of the conductor, negative charge in the conductor will flow out to pair with that positive charge. Like in the previous case, this will cause all the negative charge in the wire to move slightly to the end where the positive charge entered, allowing negative charge to flow in the other side (which is the same as positive charge flowing out of that end. I know it is weird, but it is actually true. In a charge neutral object, you cannot tell if a + charge flowed in, or a negative charge flowed out). The net result in both cases is that charge can flow through the wire, but the wire remains charge neutral: when a charge enters the wire, another charge at the other side leaves the wire.

1.2.3 Switch

A switch is an electrical/mechanical device that is used to make or break a connection between conductors. A button or lever moves a conductor so that it makes physical contact between the two terminals.

We describe the state of a switch as disconnected (off) and the charge can't flow, or as connected (on), meaning the path for charge is complete. While the switch is disconnected, electrons try to flow into the negative wire but they have no place to go, and positive charges can't flow into the positive wire either.

Question: What happens when we close the switch and connect the two wires?

When the switch is connected, the negative charges flow toward the positive terminal of the battery, pulling electrons out of that end, or you could say that positive charges from the '+' end of the battery flow through the wire to the minus end of the wire. In other words the electrostatic force from the battery drives the mobile charge in the wire to move. As we will see soon, this moving charge is called *current*.

1.2.4 Light Bulb

Incandescent light bulbs are one way to create light with electricity. In the 2010's, incandescent lights were rapidly being replaced by LEDs, which last much longer, are more robust, and use far less electrical energy for the same light output. However, incandescent lights are conceptually much simpler, so we'll continue with them for now, and talk about LEDs later in Chapter 2.

An incandescent bulb works by heating up a thin wire, called a *filament*, so much that it glows and emits light. Normally when something heats up to be very hot, it immediately burns up and breaks, so it is somewhat surprising that incandescent bulb lasted as long as they did. Making them last took lots of engineering work: the filament is made of a very durable substance (typically tungsten), and the bulb is evacuated or filled with an inert gas to prevent the filament from oxidizing.

The energy used to heat up the bulb comes from the battery, which pushes charge carriers through the bulb. The battery converts chemical energy to electrical energy. The bulb converts this electrical energy into heat (and a little light). We need to wait until Chapter 2 to explain this fully, but the conductor inside the light bulb is not a very good conductor of charge. This means when the charge tries to flow through it, the carriers are constantly "bumping" into the stationary atoms. If you continue to hit something it will get warm, since on each collision the kinetic energy of the particle is converted into heat. The amount of energy released by these collisions is proportional

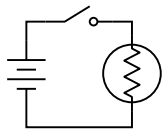


Figure 1.1: A schematic diagram of a flashlight with a symbol representing a battery (set of horizontal lines), a switch (a break in the wire that looks like you can close it), and the light bulb (a set of diagonal lines)

both to the number of collisions and to the average energy each collision releases. If there is enough collisions/second, the wire can get hot enough to glow, and it is this very high heat that emits light. That is why incandescent lights are always so hot. They have to be hot to emit light.

So a flashlight converts chemical energy in a battery into an electrical signal that a bulb converts to heat which generates the light that we see. While we could always talk at this low level with charge movement, it would be nice if we could think about things at a level that is a little more abstract. We will introduce those abstractions in the next sections.

1.3 Circuit Abstraction

Abstraction is an important part of engineering. While nearly everything we do can be broken down and analyzed in terms of fundamental principles, very often it is advantageous to create higher-level models. This is exactly what we will do here. While we could continue through the course attempting to analyze charge distributions and electrostatic forces in various devices, it is much easier to analyze circuits in terms of simple rules which encapsulate their more fundamental behavior. In the next few sections, we'll build on what we've already seen with the flashlight circuit, explaining its operation in quantifiable terms. In doing so, we'll define terms and equations that will be broadly useful as we consider other circuits.

Later, we'll see how several layers of abstraction can be stacked on top of each other, allowing engineers to understand and construct extremely complex circuits such as computer chips. We may occasionally come back to the ideas of charges and forces, to provide an alternate explanation of the operation of additional circuit devices, but for the rest of the class we will use the concepts of *current*, *voltage*, and *power*. Before introducing those concepts, we need to describe the basic nomenclature of a circuit, and its symbolic shorthand: a schematic diagram.

1.3.1 Schematic Representation

Figure 1.1 shows an electrical schematic describing the flashlight. The schematic is an abstract representation that allows us to consider the electrical behavior of the circuit independently from its physical layout.

Circuit devices are represented with various symbols, and wires are represented by lines connecting them. Figure 1.2 shows a number of the different circuit devices, also called electrical devices, that are commonly used. The behavior of all these devices, and a few additional ones will be described in later chapters. Before we talk about the properties of these devices, we first need to understand how to decode the information that is contained in the diagram.

Figure 1.2: A list of the almost all of the device symbols that we will use in Engr 40M. The properties of all these devices will be explained in later chapters.

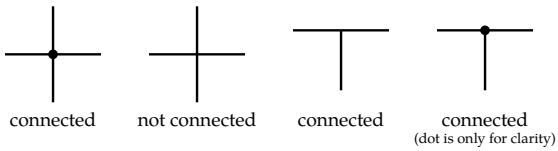
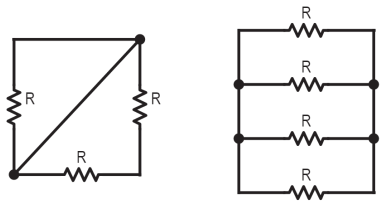


Figure 1.3: A list of possible line crossings and what they mean in a circuit schematic drawing. In general when two lines cross they are not connected unless you put a dot there to indicate you want them connected. This convention allows wires to cross without forcing them to connect.

The lines in the diagram show how the different devices are connected to each other. If there is a line connecting one end of an element to one end of another element (for example from the bottom of the battery to the bottom of the light bulb in Figure 1.1, that means that the ends of these two devices are electrically connected. In other words charge is free to flow from the battery to the bottom of the light bulb. Similarly, the line between the top of the battery and the left side of the switch indicates that charge can flow between these two devices as well. Since the lines in this diagram just represent connections between the electrical devices, each line (no matter how long and how many branches it has) is called a *node*. Every point in a schematic connected by this line is part of the same node, regardless of how the connections are arranged. Basically the line just represents that the terminals of all the devices that it touches are electrically connected together, and therefore must share some electrical properties. There is one tricky point about this diagrams. What happens when two lines cross? The standard convention is shown in Figure 1.3, and shows that crossing wires are not connected unless there is a junction marked with a dot.

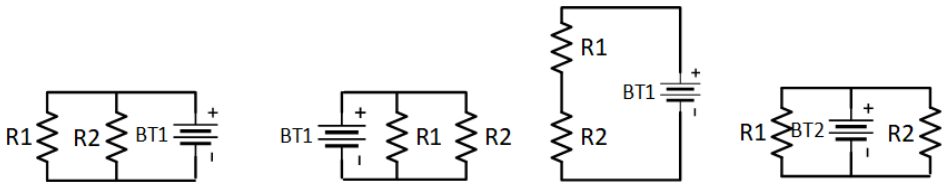
Problem 1.1
In each of the schematics below, how many nodes are in each circuit? What are they?



One surprising characteristic of schematic diagrams is that every circuit can be drawn a number of different ways! Since the lines just represent which devices are connected to each other, the order

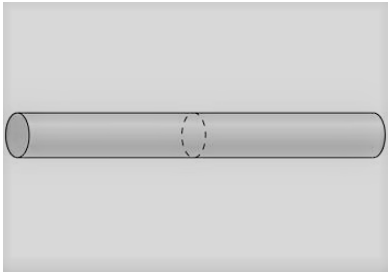
of the devices in the diagram doesn't matter. All that matters is which device connects to which. This can be useful when trying to analyze a circuit, since sometimes you can make the problem easier to understand by redrawing the schematic.

Problem 1.2
Which of the following schematic diagrams represent the same circuit?



1.4 Current

Electrical *current* is a measure of the amount of charge that is moving through a device per unit of time. Current is one of the two most important electrical quantities that we talk about. If the charges moving in the device are electrons, you can think of it equivalently as the number of electrons moving into a device (or through a plane in the device per unit time). In the figure below, you could be counting the charge that crosses the dotted cross-sectional area.



Current measures charge per unit time, so the most natural unit is Coulombs per second. This is renamed the Ampere (amp for short) in honor of Andre Ampere. [1]

For historical reasons, electrical engineers define current as the flow of positive charges. The charges which actually move in a wire are often electrons, but as we mentioned earlier, it doesn't matter whether we say negative charges move one way or that positive charges move the other since the circuit can't tell the difference, so this historical mistake is only a minor annoyance. This is called *conventional current*, and works just fine for all the circuit analysis we're going to do.

Another historical convention is that current is generally represented as 'i' and not 'c' in equations. So when you have an equation to find the current it would read, $i = f(...)$. This is because initially people talked about the intensity of the current, so the natural variable was i, but later we stopped taking about intensity.

[1] Because the SI system aims to define everything in terms of fundamental units, the ampere is actually defined as the current which produces a magnetic field causing a particular force between two straight wires of a specific length in a vacuum. But that is not particularly helpful to us here, and we'll stick with Coulombs/second.

The next couple of questions are fun facts to know, but understanding the speed that electrons move through a device is generally not important. What is important to understand is that the speed of an electrical signal flows very fast: at the speed of light.

Question: How quickly does an individual electron move through the wire?

This is actually a tricky question, because electrical signals travel much faster than the electrons. When you you add charge to one end of a wire, the electrons you add push other electrons already in the wire, which in turn push other electrons in the wire. This information wave (that electrons are being added) travels at the speed of light (very fast). But that doesn't mean the the carriers travel very rapidly. In fact, electrons travel quite slowly through wires.

We can approximate how quickly with some simple calculations. Assume we have 1 A flowing through a 22-gauge wire (which has a diameter of 0.64 mm). Each millimeter of wire contains about 2.7×10^{19} copper atoms.

An amp is about 6.2×10^{18} electrons per second flowing past a point. If each copper atom in the wire can be associated with a moving electron, then $\frac{6.2\times10^{18}}{2.7\times10^{19}} = 0.23\,\frac{\text{mm}}{\text{sec}}$.

This calculation is approximate, but it should give you a rough physical intuition about how electrons actually move in the wire. This velocity is known as the *drift speed*.

Question: When you flip on a light switch, the light turns on practically instantaneously. Why?

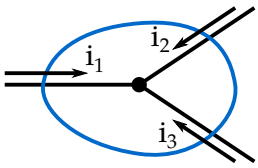
It's tempting to think that the charges flow quickly from the source (the live wire in the wall) to the light bulb, but we already saw that the individual charges actually flow quite slowly.

Instead, the light turns on quickly because charges are already in the wire, and the electrostatic force pushes them along propagates at near the speed of light. When the switch is flipped on, electrons throughout the entire wire begin to flow. Current doesn't have to start at the switch and travel to the light bulb; current begins flowing around the entire loop as soon as a complete circuit is made.

One constraint that we have is that all electrical devices and nodes (the lines connecting devices) should be charge neutral. If a current is flowing into one terminal of a device, and current is the flow of charge, the only way for the device to remain charge neutral is for the same current to flow out of the device through the other terminal. This means that the net flow into any device must always be zero.

This charge neutrality constraint means that currents must flow in a loop. If current flows out of the top of a device, the same current must flow into the bottom of the device. This is true for all the devices that the current passes through, so it must flow in a loop. If the circuit only has one loop it is very easy to analyze, since the current through all the devices that form the loop must be the same. Unfortunately circuits often have multiple loops, and in this case the current can split and flow through two different paths.

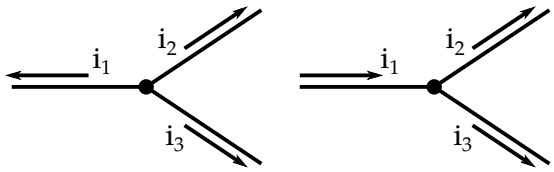
Question: Some circuits have multiple loops. What happens to current at a junction?



Let's draw a circle around the junction. Current can't pile up at the junction, which means that for every charge going into the circle, there must be a charge going out on one of the other branches. We can write this as an equation in terms of the current on each wire:

$$i_1 + i_2 + i_3 = 0$$

Of course, we might draw the circuit with the reference directions going the other way, and then we'd need to flip the signs in the equations. All the current going in must equal all the current going out. Another way to say this is that the current that is flowing into any node must be equal to the current flowing out of that node (remember that a node is what we call the lines that connect different devices).

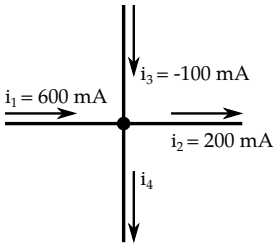


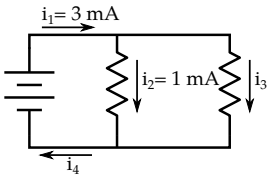
$-i_1 - i_2 - i_3 = 0$ $\quad$ $i_1 - i_2 - i_3 = 0$

$i_1 + i_2 + i_3 = 0$ $\quad$ $i_1 = i_2 + i_3$

This fact is known as *Kirchoff's current law*, abbreviated KCL and it applies to nodes and devices.

Problem 1.3
Find the missing currents:





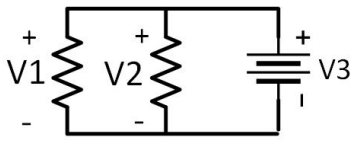


Figure 1.4: Because there is no voltage drop across a node and all three components are connected to the same two nodes, they must have the same voltage. V1=V2=V3

Question: How much is one Ampere?

For most electronic devices, a few hundred milliamps is a lot. Devices powered by AA batteries draw tens of milliamps. But some common things use quite a lot more. A cell phone uses tens to hundreds of milliamps. An 60-watt incandescent bulb draws about 500 milliamps. A desktop CPU might draw tens of amps, and the motor for an electric car may pull hundreds of amps. However, current isn't the whole story. The power delivered also depends on the voltage, which we'll talk about next.

1.5 Voltage

So far we have talked about charges and the forces that they generate on each other. Now as the chapter mentioned earlier and you might (or might not) remember from physics, it takes work (energy) to move an object when a force is acting on it. That is the reason it is harder to bike up a hill than bike on a level surface. It turns out that the energy it takes depends only on the integral of the force along the path taken. In our case, with an electron with a constant electrostatic force on it, the amount of energy - either gained in acceleration or expended in pushing - is proportional to strength of the force times the distance over which the particle is moved.

We use *voltage*, measured in Volts (Joules/Coulomb) to measure the the change in energy per unit charge caused by this electrostatic force. Voltage for charge is analogous to height for fluids. The voltage across a device is a measure of how much energy is needed (or gained) to move a unit of charge across a device, just like height is a measure of the change in potential energy for a physical object.

Since voltage measures the potential energy *difference*, is always relative and must be defined between two points. As a result, voltages are labeled between two points (usually across a device), and we talk about the "voltage across" a device (meaning the voltage difference between one terminal of the device and the other), or the "voltage between" two nodes. Since a node (the lines on a schematic) just represents the connection of the devices it touches, there is no voltage difference between the node and the terminals that it connects to. So the voltage across a device is always equal to the voltage between the two nodes that it connects to. This is shown in Figure 1.4.

As you might have guessed by now, a battery's voltage rating is a measure of the potential difference it sets up across its two terminals. The chemical reaction in the battery gives the charge at the '+' end of the battery more potential energy (like carrying the charges up a hill). If there is a path for charge to flow from the '+' to '-' terminal of the battery in the external circuit, it will (causing a current to flow) since the higher potential charge naturally wants to move to lower potential (or said differently, the electrostatic force in the battery will push charge around around the external loop).

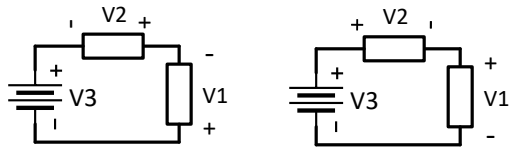


Figure 1.5: This figure shows two simple circuits (they might be the same circuit, but with different choices for voltage reference directions). The sum of the device voltages around any circuit loop must equal zero, and this constrains the values of the device voltages. To check this constraint, we will travel clockwise around the loop (you get the same result if you travel counter clockwise). For the circuit on the left, the energy of charge increases by V3 (since it travels from the - terminal to the + terminal), then increases by V2 and finally increases by V1 to return to the bottom of the battery. Thus for this circuit, $V1 + V2 + V3 = 0$. For the circuit on the right, the energy of the charge first increases by V3, but then decreases by V2 (since it travels from the + terminal to the - terminal; if V2 was positive, the charge would have lost energy going through the device) and then decreases by V1. So for the right circuit, $V3 - V2 - V1 = 0$, or $V3 = V1 + V2$

If an electron goes all the way around a loop in the circuit, it ends up at the same device terminal that it started at. In our flashlight, let's assume that the switch is connected. The chemical reaction in the battery provides energy to push a negative charge (electron) from the positive terminal to the negative terminal of the battery. This charge can then flow through the switch and light bulb, where it loses energy, and finally back to the the positive terminal of the battery where it started. Since the charge is at the same position it started at, the net energy change along the path should be zero. In other words, the amount of energy given to it by the battery must equal the amount of energy lost as it travels around through the other devices in the loop. This requirement that the sum of the energy change along any loop is analogous to the physical constraint that the net change in height for any fluid flowing around a loop is zero.

Since the energy of any charge flowing through the loop must come back to its starting energy, therefore adding all the voltage differences across the devices in the current loop must equal zero. Figure 1.5 shows two circuits that each have three devices in series. In both cases the sum of the voltage drops that the charges see flowing through the loop must equal zero, but the resulting equations depend on how the device voltages are measured. It is extremely important that you pay attention to the reference direction of the voltage measurements!

Like the current law, which is really about charge conservation, this law, which is about energy conservation, also has a fancy name. It is is known as *Kirchoff's voltage law*, or KVL for short. With KVL and KCL together, we have a very powerful way to analyze circuits.

The fact that the voltage around any loop is always zero also means that the voltage difference between any two nodes is always well defined. In Figure 1.5 the voltage from the node that connects the bottom of the battery to the node that connects to the top of the battery is $V3$, but it is also $-V2-V1$ if I run through the right half of the left circuit. By KVL we know that $V1+V2+V3 = 0$ so $V3 = -V1 - V2$ so the voltages are the same. Similarly for the right hand circuit the voltage is $V1 + V2$ which is the same as $V3$. Thus it is possible to choose one node in the circuit as the reference node, and measure the voltage difference between all the other nodes in the circuit and

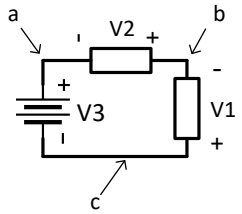


Figure 1.6: In this circuit we made node 'c' the reference. This means that the voltage of node 'c' relative to node 'c' is zero. There is no voltage difference between a node and itself. The voltage difference between node 'a' and node 'c' is V3, so we can say that that the voltage of node 'a' is V3 (relative to node 'c'), and the voltage of node 'b' is -V1 (relative to node 'c'). Using just these node voltages, $V_a = V3, V_c = 0, V_b = -V1$, one can compute all the device voltages. The voltage across the battery must be the voltage of the positive terminal minus the voltage of the negative terminal or $V_a - V_c = V3$. The voltage across device 1 must be $V_c - V_b = V1$ (remember to always start with the node connected to the positively labeled terminal). The across device 2 is $V_b - V_a = -V3 - V1$, which is the same result as was given in Figure 1.5.

this node. This is shown in Figure 1.6.

Problem 1.4
While Figure 1.6 chooses node 'c' to be the reference, the equations still work no matter which node was set to be the reference. Try setting node 'a' to be the reference. Find the other nodal voltages and confirm that the device voltages are unchanged. In this case $V_a = 0$, but V_c is no longer 0 ...

Since we have been noting that voltage for charge is similar to height for physical objects, we take one other idea from them. In mechanics, the potential energy of an object is given by $PE = m \cdot G \cdot h$, where m is the mass, G is the acceleration of gravity, and the height h has to be defined as the vertical distance between two points. Like with height, it's often helpful to designate one point as the "zero point" and measure everything relative to that. In mechanics, this is usually the ground, and in electronics, we refer to the node we choose as our reference as *ground*.

Once one has defined a ground reference, they often refer to a node's voltage, or the voltage "at" a node. This doesn't mean that a single node has a voltage, it's just a lazy shorthand for saying that there is a voltage between that node and the (ground) reference node. Often we'll pick ground to be the lowest potential of the circuit (i.e., the most negative point) so that the voltages from all the other nodes to ground are positive. However, this is not always the case.

In the example circuit shown in Figure 1.7, the voltage across the battery is 10 V, which means that node 'a' has a potential of 10 V relative to ground. Node 'b' has a potential of 2 V relative to ground, since device 3 has 2 V across it. Finally, we can use a shorthand version of KVL[2] and calculate that the voltage across device 2 must be $10\,\text{V} - 2\,\text{V} = 8\,\text{V}$.

[2]We could write out the loop equation based on the voltage across each element, and solve it. But since we already know the node voltage (relative to ground) at both ends, we can just do a simple subtraction.

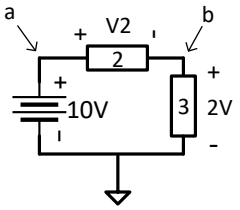


Figure 1.7: In this figure the common node has been represented by connecting it to a little triangle. This is the symbol often used to denote ground. Relative to this ground, the voltage of node 'a' is 10 V, and node 'b' has a voltage of 2 V. This means that the voltage across device 2 is 8 V.

Question: How much is a volt?

Single-cell batteries are generally between 1.2 V and 4 V, depending on the particular chemistry. Commercial electronic devices often run at 3.3 or 5 volts, although modern devices run at lower and lower voltages to save energy. The electricity in your home is at 110 or 220 volts. High-voltage transmission lines are on the order of 150 kV.

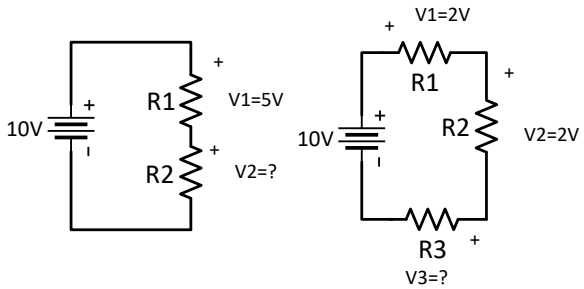
Question: Can high voltages kill you?

High voltages by themselves won't kill you. What is dangerous is the currents that high voltages can cause. For example, the shock you get on a doorknob or a car door can easily be hundreds or thousands of volts, but the amount of charge that flows is very tiny.

Your skin is quite a good insulator, so it would take hundreds or even thousands of volts to kill you. However, when your skin is wet, it conducts electricity much better, making even the 110-volt power in your home potentially lethal. For this reason, outlets in kitchens and bathrooms have special built-in circuit protection (known as GFI's or GFCI's), and appliances you might use near water (such as hair dryers) carry additional warnings.

Problem 1.5

Find the missing voltage for each of the circuits below. You should be applying KVL to do this, or you can use nodal voltage analysis.



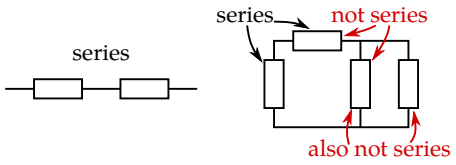


Figure 1.8: The partial circuit on the left shows two series devices. The left device connects to the right device at a node, and there are no other devices that connect to this node. The right part of the figure shows a complete circuit and which devices are in series and which are not. While there are many devices that match the first condition (connect through a middle node) only one pair of devices doesn't have an addition device that also connects to that node.

1.6 Circuit Simplification: Series and Parallel Devices

Now that we know about the rules governing current flow and voltage loops, there are few ways that you can connect devices that are special, and are easier to analyze than the general method described in Chapter 3. In this section we formally define these types of connections: series and parallel. In both cases the resulting circuit also looks like a two terminal device (just like each device inside of it), and the current and voltage across the combination of the two devices can be easily found (which is what makes them special).

Two device are in series if two conditions are meet:

1. A terminal of the first device connects to a terminal of the second device.
2. No other devices are connected to this "middle" node. That is there are only two connections to this middle node, and they are from the devices mentioned in the first item.

Figure 1.8 gives some examples of series devices, and devices that might look like they are series connected, but aren't. What makes this type of connection special is that by charge conservation, aka KCL, all the current that flows through one device must flow through the second device: every charge that leaves one has no choice but to go through the other. This is just a special case of the KCL equation that we mentioned earlier, that the sum of the currents flowing into any node must be zero. Since the current through the two devices is the same, the voltage across the series sub-circuit (the two devices in series) is just the sum of the voltages across each device (for the given current). So for series devices, if I know the current through the path it is very easy to find the voltage across the pair.

Parallel devices take advantage of energy conservation (the voltage loop law or KVL) instead of KCL. Two devices are in parallel if these two condition are met:

1. A terminal of the first device connects to a terminal of the second device (the same rule as a series connection).
2. The other terminal of the first device connects the other terminal of the same second device.

These constrains say that two devices are in parallel if they share both nodes, i.e., they are connected on both ends. Unlike series connection, for parallel connections, it doesn't matter how many other devices are connected at these nodes, all that matters is that both devices share both nodes. Figure 1.9 give some examples of parallel devices, and devices that might look parallel

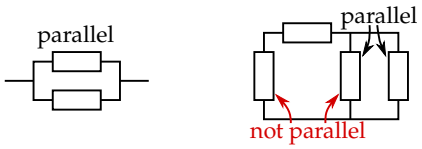
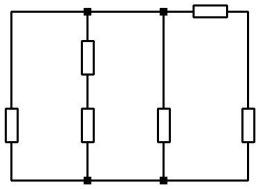


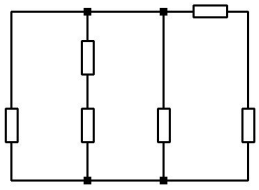
Figure 1.9: The partial circuit on the left shows two parallel devices. Both terminals of the top device connect to the corresponding terminals of the bottom device. The right part of the figure shows a complete circuit and which devices are parallel in this circuit. For parallel devices, it doesn't matter if other devices connect to the shared nodes. All that matters is that both device terminals are connected to each other.

but aren't. What makes this connection special is that we know from energy conservation (KVL) that the voltage across both devices must be the same: since voltage is the difference in potential between two nodes, and the devices share the same two nodes, the two device voltages must be the same.[3] Thus for parallel devices the total current that flows through the combination is the sum of the currents through each device.

Problem 1.6
Identify which of the devices in the following circuit are in series. (For simplicity, only consider a combination of a maximum of two devices at a time)



Problem 1.7
Identify which of the devices in the following circuit are in parallel. (For simplicity, only consider a combination of a maximum of two devices at a time)



We use series and parallel circuits when we want to measure a device voltage, or the current running through a device. To measure a voltage across a device, the meter needs to be in parallel

[3]Be careful about the reference directions for the two devices. If they are labeled differently, for example one has '+' on the top, and the other has '+' on the bottom, then the measured values will be opposite signed.

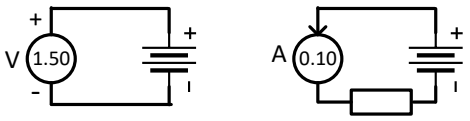


Figure 1.10: *Left circuit:* To measure the voltage across the battery, you put your measurement device in parallel with the battery. This forces the voltage across the measurement device to be the same as the battery voltage. *Right circuit:* When you want to measure current, you need the current through the measurement device to be the same as the current you want to measure, so you need to place the meter in series with the device you want to measure. The right circuit measures the current that is flowing out of the battery.

with that device. Since the meter is in parallel with the device, it will have the same voltage across it as the device, so the voltage it measures is also the voltage across the device. To measure current, the meter must be in series with the element, so that the current passing through (and measured by) the meter is the same as the current through the element. These connections are shown in Figure 1.10

1.7 Power

Now we understand what a circuit schematic is, and the rules that govern current flow and voltages in a circuit, we have all the basic tools needed to start analyzing how circuits behave. But the reason we started looking at all this material was to try to understand how a flash light or a solar charger works, and both of them use electrical signals to move energy around. So before ending this chapter, this section looks at how to analyze energy flow in circuits.

We now have the basic concepts to understand how electrical signals can move energy around in a circuit. Some devices give the moving charges (the current) extra energy by raising the potential energy of the charges that flow through them. A battery is such a device. This extra energy is then delivered (lost) in some other device, or devices, as the current flows in a loop. It must be lost, since when the charge enters the lower voltage terminal of the battery, it needs to have the same energy it started with.

It is a common misconception to assume that the current somehow gets "used up" in the circuit. This is incorrect, because the charges in the circuit don't get used up or destroyed: remember that the charge in every device is conserved! However, the charges flowing in the circuit do gain and lose potential energy, and these changes are what move energy around in the circuit. To help see this, it may be helpful to think about the fluid analogy again. The battery is like a pump, pumping water up and providing potential energy. This energy can then be used to turn a waterwheel (or a hydroelectric dam). In these devices the water falls from a high place to a low place - and does work in the process - but the water does not get used up, it just flows in a loop like the charge[4].

If voltage across a device represents the change in energy per unit charge, and current is the flow of charge per second, then voltage times current, $i \times V$, is the energy/sec that the charge gains

[4]And unlike water which can evaporate, the charge doesn't leak away by other means

(or loses) flowing through the device. This energy change is measured in Joules per second (aka Watts). Energy/sec is just power, so this product represents either the power that this device is supplying to flowing charge (and the rest of the circuit), or the power this device is using.

At this point, we need a convention to keep track of whether an element is dissipating power (like a light bulb) or supplying it (like a battery). To figure this out is not hard, if you remember what voltage and current mean. Take your device, and figure out which terminal has the higher voltage. If the measured device voltage is positive, it is the terminal labeled '+'. If the measured voltage is negative, this means that the voltage on the node labeled '+' is actually lower than the node labeled '-', so the node labeled '-' is actually the higher terminal.

Once you have the node which truly has the higher voltage on it, look to see if current is flowing into the device from this higher voltage terminal. If the current is flowing **in** to the device, then the charge is flowing from the higher voltage to the lower voltage, so each charge is losing energy, and this device is pulling energy from the circuit. If on the other hand, the current is flowing **out** of the higher voltage, it is adding energy to all the charge that flow through it, and this device is supplying energy to the circuit.

What is interesting is that you actually don't need to find the terminal with the higher voltage. If you simply multiply the measured voltage (whether it is positive or negative) by the current that flows into the terminal that is labeled '+' you will get the same result (you should work out the algebra to prove it to yourself. This is called the *passive sign convention*.

Passive sign convention in practice

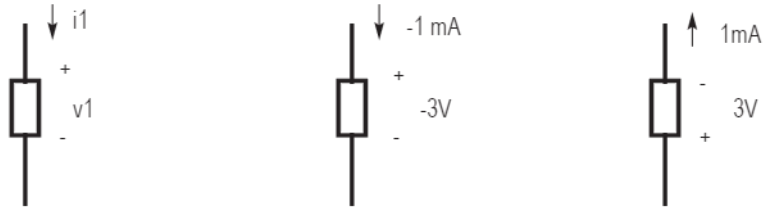


Figure 1.11: Sign Conventions

Consider Figure 1.11. Let's say we have a two terminal device, and we measure the voltage across it and the current through it in the way shown. We don't know which terminal actually has the higher voltage, or which way the current is flowing, and arbitrarily measure and label the voltage/current as shown in the left-most image. By our convention, the power of this device is $v_1 \times i_1$.

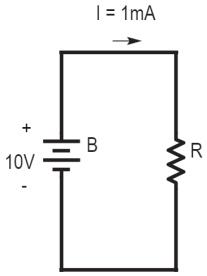
Our measurements under this setup turn out to be $-1mA$ and $-3V$ as shown in the middle image. $P = v_1 \times i_1 = -3V \times -1mA = 3mW$.

This circuit can also be redrawn as shown in the right-most image, which represents the exact same circuit. Notice how we flipped the voltage polarity and current direction, so now their quantities are positive. In this instance, $P = v_1 \times i_1 = 3V \times 1mA = 3mW$! We get the same answer. The only and key constraint is the passive sign convention. The positive direction of current **has to** flow into the terminal that you labeled at positive.

By following the convention, we are guaranteed to arrive at the same result.

Under this convention, positive values for power indicate that the device is dissipating power, and negative values imply that it supplies power to the circuit. The sum of the power dissipated by all the devices in the circuit must be zero, which is just a way of saying that all of the power must come from somewhere and all the power supplied has to get used up somewhere.

Problem 1.8
1. How much power is being supplied or dissipated by the battery on the left side of the circuit? Is this power positive or negative by our convention, and thus is the power being supplied or dissipated?
2. How much power is being supplied or dissipated by the resistor on the right side of the circuit? Is this power positive or negative by our convention, and thus is the power being supplied or dissipated?
3. What is the sum of all the power in this circuit?

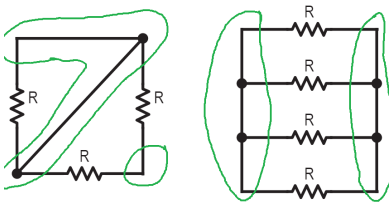


1.8 Summary

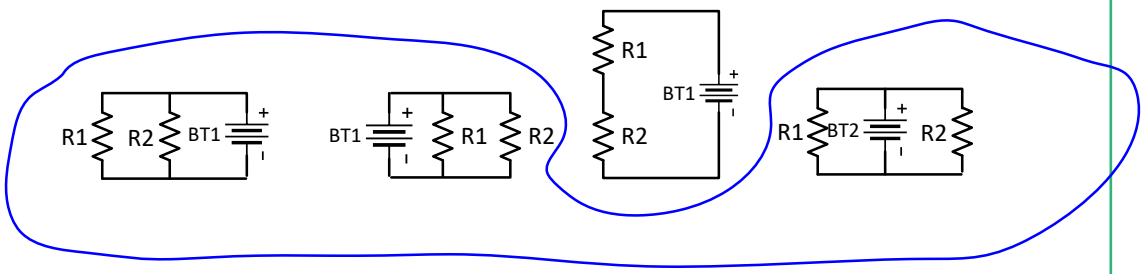
Many of the objects you use depend on the behavior of electrical circuits to operate correctly: if it has a battery, or plugs into the wall, it is a type of electrical circuit. This chapter has shown how you can represent any of these circuits using a circuit schematic diagram, and how to interpret these diagrams. It also introduced current, which is the flow of charge, and voltage, which is a measure of the potential energy difference of charge at one node from charge at a different node. Then it showed how you can use charge conservation (KCL) and energy conservation (KVL) to analyze these circuits. Finally, it also showed that you can track the energy flow in a circuit by looking at the product $i \times V$, since this product represents the power the device is absorbing from or delivering to the rest of the circuit. We almost have enough to work with circuits, but to do so requires us to understand the relationship between current and voltage for each of the different types of electrical devices. We tackle this issue in the next chapter.

1.9 Solutions to practice questions

Solution 1.1:
Each circle has only two nodes, which are grouped as shown below.



Solution 1.2:
The circled circuits represent the same circuit. Three of the four circuits are the same.

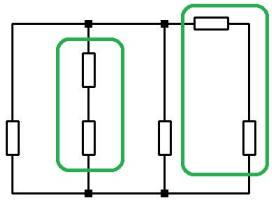


Solution 1.3:
1. $i_4 = 300mA$
2. $i_3 = 2mA, i_4 = 3mA$

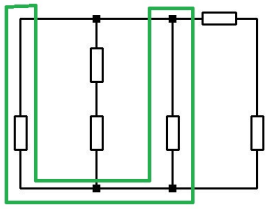
Solution 1.4:
If node 'a' of Figure 1.6 is chosen as the reference, then the voltage difference of this node relative to itself is clearly 0V. The voltage of node 'c' is now V3 volts lower than the reference, or -V3. The voltage of node 'b' is either V2, or -V3-V1. Both are correct answers, since we know that V2= -V3-V1 from KVL.

Solution 1.5:
A. $V_2 = 5V$
B. $V_3 = 6V$

Solution 1.6:
The following devices are in series.



Solution 1.7:
The following devices are in parallel.



Solution 1.8:
In the figure the current flowing into the battery is -1mA, since the reference direction for current flow is into the '+' terminal and the current is shown flowing out of the '+' terminal. Thus the battery is dissipating 10V * -1mA = -10mW. It is providing 10mW of power to the circuit. The right device is a resistor and it is in parallel with the battery, so we know it also has 10V across it, if we label the voltage the same way as the battery. With this labeling the current is also positive at 1mA, and the power is 10V * 1mA = 10mW. Since this is positive, the resistor is dissipating this energy. Since we know that energy is conserved, adding the energy of all the devices better equal zero, as it does.

Note: I could have placed the positive label for the resistor (the right device) on the bottom. If I did that, then the voltage across the device would be -10V, but positive current would flow in from the bottom and out the top, so the current would be negative as well. With this reference direction, I would get -10V * -1mA = 10mW, as we promised it would.

Chapter 2

Electrical Devices

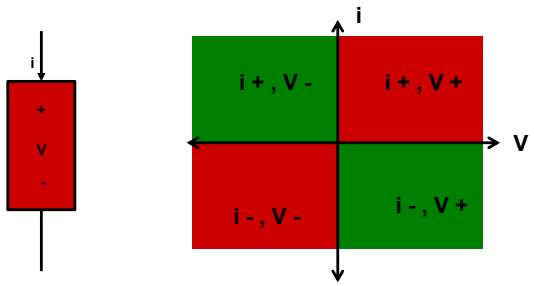


Figure 2.1: The reference directions and axis for measuring the iV relationship of an electronic device. The voltage across the device is plotted horizontally, and the current through the device is plotted vertically. Notice that the voltage and current can be positive or negative. Using the standard reference direction, which set positive current to be current flowing into the + terminal and out of the - terminal, the quadrant indicates whether the device absorbs energy (red) or provides energy (green).

The previous chapter introduced the core ideas behind what makes something electrical. Electrical circuits are connections of electrical devices, and electrical devices are things that allow charge to flow through them, where the charge is driven by voltage differences, and we measure the flow of charge as a current. The rules, KVL (energy conservation) and KCL (charge conservation), allow us to reason about the voltages and currents. But to actually understand how a circuit functions we need one more piece of information: the relationship between the current flowing through a device and the voltage across it. This chapter provides the needed relationships. Since many different devices' current-voltage relationships have similar forms, we break all devices into different classes. Devices in each class have the same type of current-voltage relationship. In this chapter, we introduce the first four classes of electrical devices: voltage and current sources, resistors and diodes.

Since we will characterize any electrical device by how the device current depends on the voltage across the device (or conversely how the voltage depends on the current), we need a good way to represent this relationship. It is most easily visualized as a plot of the device current vs. the voltage

across the device, and so we will use this plot to help us understand a device's behavior/operation. Figure 2.1 shows a generic device with our standard reference direction defined for voltage and current (positive current flow into the terminal labeled positive). Remember that the labels just indicate how to interpret the value, and doesn't mean that the voltage across the device measured from the '+' terminal to the '-' terminal is going to be positive. Since the labels are only the reference directions, we need to know how the device operates when the voltage across it is positive or negative using these reference directions, and how the device operates with current flowing in the reference direction or opposite to (negative) the reference direction. Thus our plot has four quadrants, one for each possible combination of positive and negative voltages and currents.

If we think back to how we measure power, iV, we can see that the quadrant that we operate in determines whether the power is positive (shown in red), or negative (shown in green). When the power is positive, the charges flowing in the current lose potential energy flowing through the device. Either the charge flows in to the higher voltage end of a device ('+' V '+' i case), or current is flowing out of the terminal labeled '+' but because the voltage is negative, this is actually the lower voltage side ('-' V '-'i case).[1] Since the charge carriers are losing energy, the device must be absorbing this energy. Conversely, if the iV is negative (the green regions) current is flowing into the lower voltage terminal and out of the higher voltage terminal, so the device is supplying energy to the circuit (it is giving the charge carriers more energy). Since physics says that energy must be conserved, unless a device has an internal source of energy, it must operate only in the red regions.

With this way to visualize current voltage relationships, the rest of the chapter will discuss a few of the important device classes.

2.1 Voltage Source

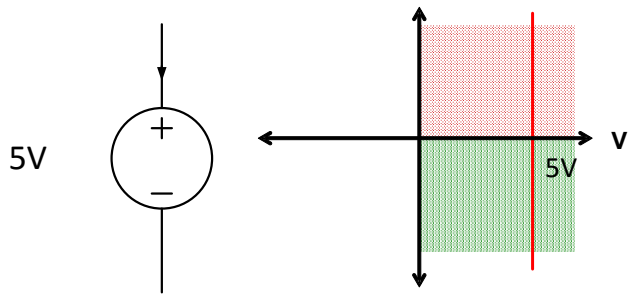


Figure 2.2: A voltage source. On the left is the symbol used for a voltage source, and on the right is the iV curve of the device. Since the voltage across a voltage source is constant, the resulting curve is a vertical line — it can support any current needed to maintain the voltage across the device.

We have already been introduced to a couple of types of devices in the first chapter. We described a battery as a device that tries to keep the voltage across it constant. The device class with this type of characteristic is called a *voltage source*. The voltage across an ideal voltage source is always

[1]Note that this case also means positive current is flowing into the terminal with the higher potential, since if current is flowing out of the lower voltage terminal, current must be flowing into the higher voltage terminal to maintain charge neutrality in the device

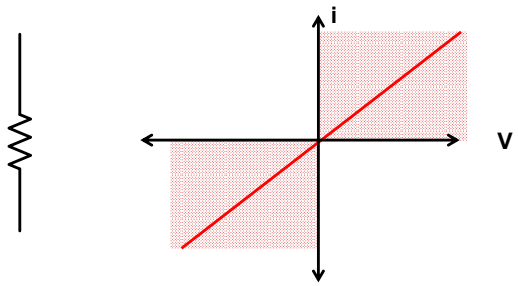


Figure 2.3: A resistor. On the left is the symbol used for a resistor, and on the right is the iV curve of the device. In a resistor, the voltage is proportional to the current, so the iV curve is a diagonal line. The slope of the line is one over the resistance $(1/R)$. Note that resistors only absorb energy, which makes sense, since they don't contain any energy sources.

constant, and equal the "voltage" of the voltage source. Since a voltage source only sets the voltage across it, it will support any current that the other devices in the circuit need to maintain its voltage. This gives the iV curve shown in Figure 2.2. The red line in this figure is the set of voltage - current points that the device supports. The line is vertical, since the voltage is always fixed independent of the current that the device supplies.

Question: Is a battery always a power source in the circuit?

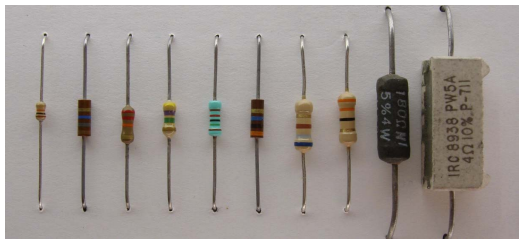
Notice that the output of a voltage source exists in both a red and green region. This is because it can either absorb or supply energy. If the voltage is positive, when current flows into the positive terminal the carriers will leave the negative terminal with less energy so the voltage source absorbs energy from the circuit. However when the current flows out of the positive terminal, the voltage source is supplying energy to the circuit.

These dual roles happen in the real world as well. Consider a rechargeable battery. When you use the battery to power your circuit, current flows from the positive terminal of the battery through your circuit (perhaps to the light in your flashlight) back into the negative battery terminal. During this operation, the battery is converting chemical energy into electrical energy which is then consumed by the circuit. But when you charge this battery, you take another source of energy (the wall socket), and current flows from the charger into the positive terminal of the battery. The battery absorbs this energy and stores it as chemical energy so it can be used again when it is needed.

2.2 Resistor

One of the most common devices that we will work with are *resistors*. A resistor is another two terminal device, but in a resistor the voltage across the device is proportional to the current running through it. In fact the proportionality constant is called the resistance of the resistor, represented by R, and it is measured in ohms, Ω. The proportionality can be written as follows:

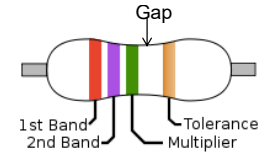
$$V = i \cdot R$$



http://ecee.colorado.edu/~mathys/ecen1400/labs/resistors.html

http://www.instructables.com/id/Reading-Surface-Mount-Resistor-codes/

Figure 2.4: Resistors come in many different shapes and sizes. The resistors on the lower right are surface-mount resistors, which are used on printed circuit boards.



- Measure a resistor
  - Does it match the value printed on it?

Color	First Digit	Second Digit	Third Digit 1% Resistors	Multiplier	Tolerance
Black	0	0	0	1	
Brown	1	1	1	10	±1%
Red	2	2	2	100	±2%
Orange	3	3	3	1000 (=1k)	
Yellow	4	4	4	10k	
Green	5	5	5	100k	
Blue	6	6	6	1000k (=1M)	
Violet	7	7	7	10M	
Gray	8	8	8	100M	
White	9	9	9	1000M (=1G)	
Gold				0.1	±5%
Silver				0.01	±10%

Figure 2.5: Resistor color code. On many resistors, the three numbers are represented by colors and not by numerals, which makes reading resistors even more fun.

The iV curve of a resistor is shown in Figure 2.3. The slope of the line is 1/R. Resistors are devices that conduct current, but it takes energy to get the charge to move. The more current you want through the device, the harder you need to drive the charge, and the more voltage will "drop" across the resistor. We will find that these type of devices are useful for building different types of circuits, so people have created materials which can create a wide range of resistance.

Since charge flowing into a resistor has more potential energy than it leaves with (aka positive current flows into the terminal with a higher voltage), the resistor absorbs energy from the circuit. We know that the power a device absorbs is just $i \cdot V$ and that the iV relationship for a resistor is $V = i \cdot R$, so the power a resistor dissipates is:

$$Power = i \cdot V = V^2/R = i^2 R$$

The power a resistor dissipates is proportional to the square of the voltage across it. While this might at first seem surprising, it make sense. Both the current through and the voltage across a resistor increase when the voltage (or current) increases, so a quadratic dependence should be expected. Where does this energy go? It gets converted into heat, and makes the resistor warm, so the heat can flow out. As a result be careful if you touch a resistor you are putting a lot of power into. It might be very hot. In fact, it can get so hot it burns up.

It turns out that almost all material, even good electrical conductors have some resistance.[2] Thus, all real wires usually have a small voltage drop across them when current flows through them. In most designs the wires are chosen so this drop is small enough so it can be ignored. Wire resistance is a problem for large structures as well as small ones. Every wonder why the wires connecting to your car battery are so large? It is because they need to supply a large current to start the motor, and thus need to have a low resistance. Similarly the wires that deliver power (110V) through your house also need to be wide to support the large currents they support. In fact, there are rules for the wire size required for different current ratings, and these wires must be protected by a circuit breaker to guarantee that the current through the wire can't exceed its rating. This is to prevent the wires in your house from getting too hot and starting a fire.

Of course in special cases, you do want the resistor to get hot. If the resistor is in an incandescent light bulb, it will get hot enough to glow and emit light. This is the reason that we represented a light bulb as a resistor in the first chapter. You also use resistors to generate heat if you have an electric oven, toaster, hair dryer, etc. The heating elements in these devices are all resistors that convert electrical energy to heat.

While real wires have resistance, the lines in our circuit schematics do not. These lines are perfect conductors, so there is no voltage difference between any points on this line. If you want to model the resistance of a real wire, you need to represent that resistance by explicitly adding a resistor to you circuit diagram. As mentioned earlier, for most circuits, the resistance of the wire is so small it can be ignored.

The range of resistance we use in our circuits is quite large. Recalling that the inverse of the resistance $(1/R)$ is the slope of the iV characteristic curve of a resistor, these large ranges in resistance will result in different-looking iV curves. Some circuits will use resistors larger than $10\,\mathrm{M}\Omega$, which is nearly a horizontal line near zero, to model the small amount of current that flows

[2] All material, except for superconductors which are a macro-scale manifestation of a quantum phenomenon, have resistance. These no-resistance materials, which only exist at cold temperatures, seem magical if you are used to normal conductors with loss. They have interesting properties, like current can flow in a loop forever without any voltage driving them. Since currents cause magnetic fields, superconducting circulating currents are used in MRI machines to make large magnetic fields, and they can even levitate objects.

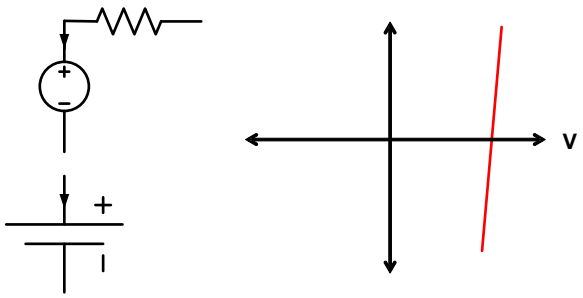


Figure 2.6: A battery. On the bottom left is the circuit symbol for a battery, and on the right is a more accurate iV plot of a battery. While the line is pretty vertical, there is some slope to it. We can model this real battery by combining a perfect voltage source with a resistor. The output voltage of this combination is going to be equal to the sum of the voltage drop across the voltage source and the resistor. When current is flowing out of the voltage source, it will flow into the resistor, so the charges will leave with less energy (voltage) then they started with, so the voltage across the combination decreases as the current becomes more negative (the current reference direction is into the positive terminal).

through a device which doesn't take much current (like a good voltage meter). Other circuits might need resistors less than 10 mΩ which is a nearly vertical line, to measure the resistance of a thick wire carrying a large current (like battery jumper cables). The difference of resistance between these two cases is a factor of 10^9.

Figure 2.4 shows a picture of different types of resistors. The smaller resistors can handle less power before they burn up (literally). The resistors on the lower right don't have leads and mount to the surface of a circuit board. They are much smaller than the other resistors that are shown. All resistors mark their resistance on them, but the marking is a little cryptic. Each resistor generally has 3 or 4 numbers on them. For example one resistor in the picture says 391. The first two numbers are put together, in this case to form 39, and this number is then multiplied by 10 raised the power of the third number (which was done to cover a wide resistance range). So 391 represents a resistor of 390 Ω. The 270 resistor represents 27 Ω.

To make reading resistor values even more challenging, on most resistors with leads, the value of the three digits are encoded by a color, as shown in Figure 2.5. This would not be so bad, but sometimes the colors are hard to figure out. Violet, Green and Grey sometimes don't look much different. When in doubt, you should pull out your trusty DMM and measure the resistance.

One final note about resistors. Just because the resistor says it is 10 kΩ doesn't mean that the resistance of the device will be exactly 10 kΩ. Most of the resistors in the lab are 5% resistors, which means that their value is within 5% of the value specified. This means the resistance can be between 9.5 kΩ and 10.5 kΩ.

2.3 Battery

We previously mentioned that a battery could be represented as a voltage source. This is basically correct. However, if you measured the output voltage of a battery as you increased the current you pulled out of it, you would notice that it would not be a vertical line. As the current you pulled

out of the battery increased, the voltage across the battery would decrease as shown in Figure 2.6. The slope of this line depends on the type of battery you have. A 9 V battery has a large change in voltage with current, whereas your 12 V car battery can supply 100 A with only a small change in voltage. Whatever the voltage drop, we can model change in voltage with current by combining our voltage source and resistor models, which is shown in the top left of Figure 2.6. The resistance of the car battery might be less than 10 mΩ, whereas a 9 V battery might have a series resistance of 100 Ω.

2.4 Current Source

Another class of devices is called a *current source*. Like a voltage source, this device only constrains one aspect of its current-voltage relationship, but unlike a voltage source which sets the voltage across it, a current source forces the current flowing through it to be equal to its set value. An ideal current source will set whatever voltage is needed on its pins to maintain this current. The symbol for a current source and its iV curve is shown in Figure 2.7

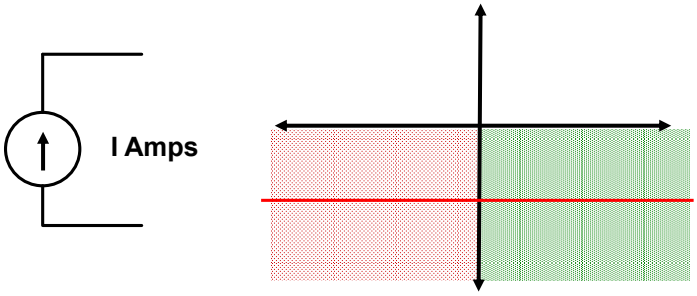
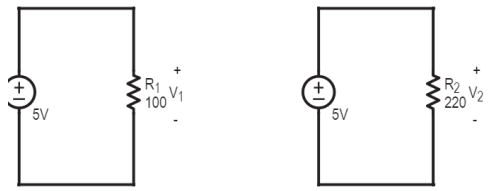


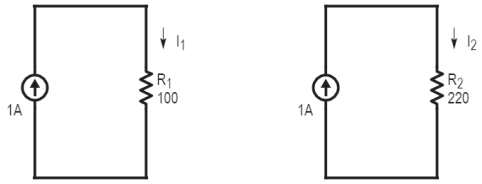
Figure 2.7: A current source. On the left is the circuit symbol for a current source, and on the right is its iV curve. If we use our standard reference directions and label the top of the current source as the '+' side, the current is flowing out of this terminal, which is opposite to the reference direction for current. That is the reason the current is negative in the plot. Since this device keeps its current constant independent of its voltage, its curve is a horizontal line. Like a voltage source, this means that a current source can either supply or absorb energy from the circuit.

While you have probably bought components that are a good approximation of a voltage source, such as batteries, you probably have not run across a current source. In fact, their behavior might just seem wrong to you. How can its voltage not be defined? However, they are very useful circuit elements, and you can buy them (or build them from other elements). Even the power supplies in the lab can be set to be either a voltage or current source. As we will see next, they also can be used in combination with other device models to model real-world effects.

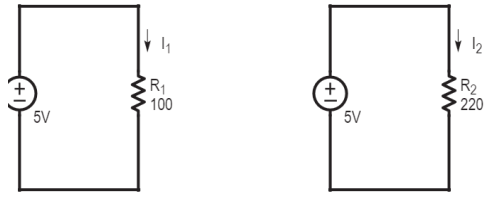
Problem 2.1 : **Voltage sources**
Find V_1 and V_2.



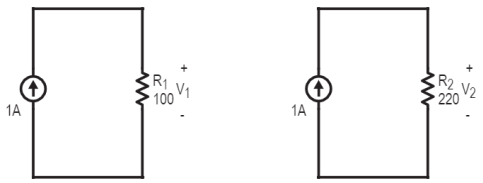
Problem 2.2 : **Current sources**
Find I_1 and I_2.



Problem 2.3 : **Simple applications of Ohm's law**
Find I_1 and I_2.



Find V_1 and V_2.



2.5 Diodes

Another electrical device that we often use is a diode. They are one of the main components in solar chargers. A diode is like a one-way valve for current, not unlike a plumbing check-valve which allows water to only flow in one direction. Like resistors, diodes don't contain any power sources, so its iV curve can only exist in the "red" regions of the iV curve. They are different than a resistor because their behavior in the two quadrants they operate in are very different. When there is a positive voltage across the diode, positive current flows, but when the voltage reverses, no negative current flows. This combination means current can flow only in one direction. This one-way behavior is very helpful in a number of circuits where you want current to only flow in one direction. For example, we would like current from the charger to flow into our phone or computer's battery to charge it up, but we don't want current from our battery to flow out into the charger (for instance if the charger is not plugged in). Diode are also useful in circuits where the voltages change over time, and we'll explore this in more detail later.

But that is not the reason we are introducing diodes to you early in this class. We need to talk about diodes because they are able to convert electrical energy to light. These diodes are called light emitting diodes or LEDs, (or similarily organic LEDs, called OLEDs, which might be in your phone screen). Diodes can also take light and convert it to electrical power. Diodes of this type are called solar cells. And we will use a solar cell (a.k.a. a diode) in our first project. The magical way they convert between light and electricity depends on some quantum physics which we will mention, but not talk about deeply.

A diode is represented in a schematic with the symbol shown in Figure 2.8. The arrow points in the direction that current can flow. Usually, the diode package will have an alternate-colored band (often silver or white on a black package) indicating the negative end (current output) of the diode which is also shown in this figure. Since current can only flow one way in a diode, we always label the voltage and current across it the same way. The positive end of the diode (sometimes called its anode) is on the side where current can flow into the device (the triangle side), and the negative end of the diode (sometimes called the cathode) is where the current can flow out of the device (the side with the line). It turns out that there are several types of diodes with special characteristics, and these have slightly different schematic symbols. These symbols are shown in Figure 2.9. We won't discuss them in more detail here, but be aware that you might see these symbols, and that they represent special-purpose diodes.

The simplest way to mathematically model a diode in a circuit it to split its behavior into two states: "on" and "off". When the diode is on, it acts like a voltage source: there is small voltage

Figure 2.8: The left figure gives the circuit schematic symbol for a diode, and the right figure shows a diode package, and how the negative terminal is usually marked with a band. Current can only flow through a diode in the direction of the arrow

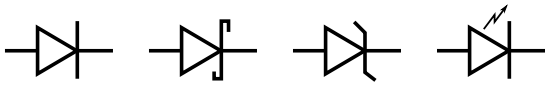


Figure 2.9: Symbols for different types of diodes. From left to right: normal diode, $V_F \approx 0.6V$, a Schottky diode, $V_F \approx 0.3V$, a Zener diode (conducts current when the reverse voltage gets to large), and a light emitting diode, $V_F \approx 2 - 3V$

.

across the terminals, and the diode can support any positive current that is needed (just like a voltage source). The voltage of this voltage source is called the *forward voltage* of the diode, and is usually represented by V_F. When the diode is off, it acts like an open circuit: zero current flows, and the voltage across the terminals is less than V_F. This is called the *idealized model* of a diode because it describes a simplified version of a diode's behavior. Needless to say, real diodes are not ideal, but this model is close to a real device and is very useful.

The idealized diode model produces a nice approximation to a real diodes iV relationship. A more correct model based on the underlying physics is given by the equation

$$I = I_S \left(e^{V_D/V_T} - 1 \right)$$

where I is the current through the diode, I_S is a scale factor, V_D is the voltage across the diode, and V_T is a constant which is approximately 26 mV at room temperature.[3] Figure 2.10 plots this exponential model of a diode on top of our idealized diode model. From the plot, one can see that the fit is acceptable.

Dealing with the diode's exponential i - V relationship is painful unless you are working with a computer program which can model these devices. Such programs exist, and we will introduce them later on in these notes. Our idealized diode model provides a good fit to a diode while still being analyzable without computer support.

Dealing with diodes is still a little complicated. To find the current or voltage through a diode, you first need to decide what state the diode is in. Is it on, or off? And to decide what state you are in, you need to know the diode's current or voltage. While at first it seems you are in a circular reasoning trap, there is an easy way out. Guess the diode's state. Once you know the state, you can solve for the diode's current or voltage and check to see if you guessed correctly. Thus to solve a circuit containing idealized diodes:

[3]Of course, this is still an approximation which assumes that the underlying physical behavior is ideal. Real diodes have resistance in series with the "ideal" diode, so the current can't exponentially increase forever, and all diodes will "breakdown" and conduct when the reverse voltage is negative enough. But for this class we will ignore these effects.

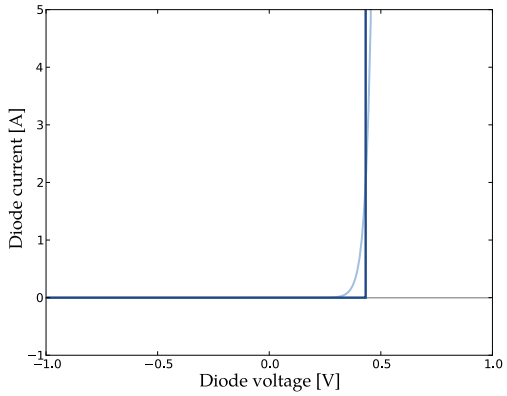
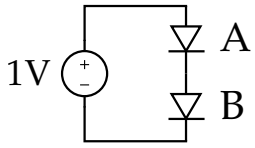


Figure 2.10: A current voltage plot showing our idealized diode model on top of the exponential current voltage model of the diode. While the fit is not perfect, especially at low positive currents, the overall fit is quite good. And being able to model a diode as either nothing (an open circuit) or a voltage source will makes analysis easier than dealing with exponential iV relationships.

- Assume that current through the diode is zero (i.e., that the diode is off, and acts like an open circuit). Find the voltage across the diode under this assumption.
- If the diode voltage is less than or equal to V_F of the diode, then the diode is in fact off, and our analysis holds.
- If the diode voltage is greater than V_F, then the assumption was incorrect, and the diode is on. The diode voltage can't actually be larger than V_F, so re-solve the circuit assuming that the diode is on and replace it with a voltage source of V_F and solve the circuit again. This time you should find positive current flowing through the voltage source, and you have solved for the current through the diode. Once you know the diode is on, you know the voltage across it. It is just V_F.

When we have multiple diodes together, it's important to consider the circuit as a whole, using KVL and KCL.

Question: In the circuit below, which diodes are on? Use the idealized model, and assume a forward voltage of 0.7 V.



If we consider only diode A, we might assume that the diode is on because the 1 V source is greater than the diode's threshold voltage. Then we go to diode B, and observe that the

voltage across it is

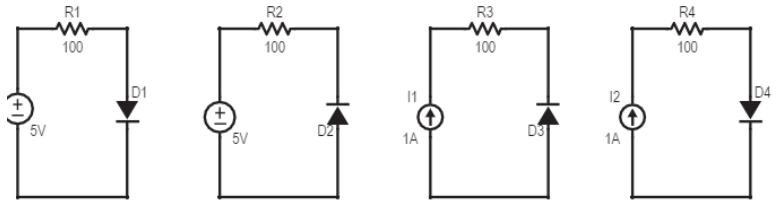
$$1\,\mathrm{V} - 0.7\,\mathrm{V} = 0.3\,\mathrm{V}$$

This is less than its V_F, so it must be off.

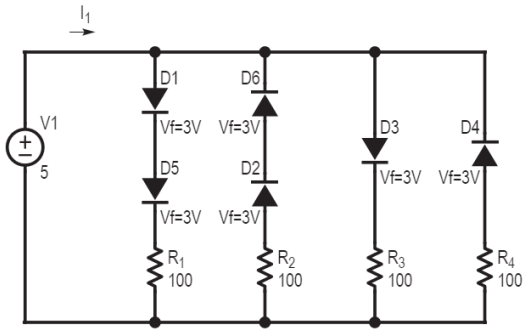
However, this can't possibly happen. If the first diode is on, current is flowing through it, and since the second diode is in series with it the two diodes must have the same current flow. So if the first diode is on, the second diode will also be on. Conversely, if the second diode is off, the first one must also be off. This is in fact what happens. If the supply voltage is not sufficient to turn on both diodes, they will both be off.

The next chapter describes the general method of solving for voltages and current in circuits, and will have some example circuits with diodes.

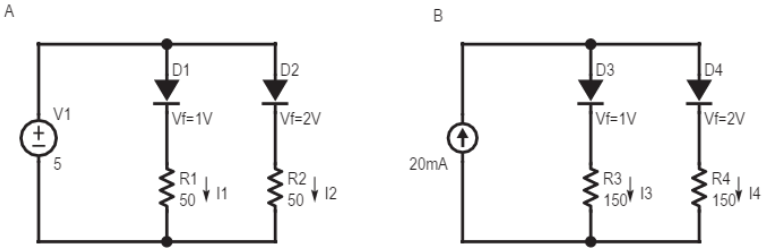
Problem 2.4 : **Zero Voltage Diodes**
Assume that all diodes in the following circuits are idealized diodes with $V_F = 0V$. Determine which of the diodes are on.



Problem 2.5 : **More Diodes**
Again, the diodes in the following circuit have forward voltages as labelled. Determine which diodes are turned on, and what the current I_1 is.



Problem 2.6 : **Even More Diodes**
The diodes in the following circuit have a forward voltage as labelled. ($V_F = 2V$ indicates a forward voltage of 2V). Determine which diodes are turned on, and what the currents I_1, I_2, I_3 and I_4 are. *Hint: Try to follow the steps in the text, especially for circuit B, which is a slight extension problem.*



2.6 LEDs and Solar Cells

One of the most interesting and useful properties of diodes is their ability to convert light power into electrical power, and convert electrical power into light. Not all diodes have this ability. Diodes that convert light to electrical power are typically called solar cells, or photo-diodes, and diodes that convert electrical energy into light are called light-emitting-diodes, or LEDs for short. Understanding the full mechanism which enable this behavior requires learning more about solid-state physics and semiconductor devices. While these are fascinating subjects, they are a little advanced for this class. The next two sections provides a high-level explanation of the physics behind diodes and light to give you a little feeling for why light can interact with diodes. If you are interested in this subject, please talk with one of the instructors of the class.

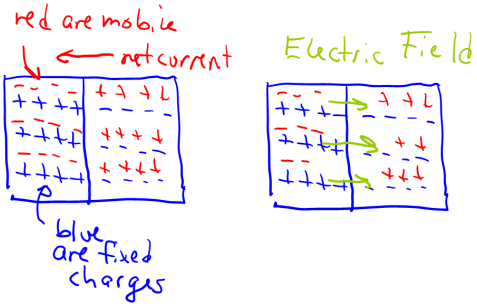


Figure 2.11: Picture of how a diode is formed. A material with mobile '-' charge is connected to a material with mobile '+' charge. Since at room temperature the mobile charges move in random directions, some of the mobile '-' charge on the left will randomly move to the right. Since there is not mobile '-' charge on the right to move left, there is a net movement of '-' charge to the right. A similar situation happens and some mobile '+' charge will move left. As a result of the charge movement, there is a charge separation in the device (it is still overall charge neutral) and an electrostatic force builds inside the device..

2.6.1 The physics behind diodes (deep background)

A diode is formed when you connect the right two types of material together. Both materials conduct charge and are, of course, charge neutral, but on one side the positive charge is mobile, and the negative charge is fixed, while on the other side the negative charge is mobile and the positive charge is fixed.[4] Mobile charges will have the ability to moves around. Now when there is not a voltage applied to it, it moves in a random directions (you can think of the charges jiggling around some moving left and some right), so there is no net charge motion, and thus no current flow. When the two sides are put together some interesting stuff happens. Lets assume that the mobile + charges are on the right and the mobile - charges are on the left as shown in Figure 2.11.

At the boundary between the mobile + and - charges, when the mobile charges begin to move, some mobile + charges move left, but there are no mobile + on the left to move right. Instead, the mobile - charges on the left move right. This gives rise to a net current to the left! This is called the diffusion current, since it is the result of the random thermal motion of charged particles. Like perfume in air, random motion will tend to spread out, or diffuse, any material with a concentration gradient. This is true whether the particles are charged or not. However, when the particles have charge, some other interesting stuff happens. When mobile + and - charge meet, they neutralize each other[5] and disappear. When charged particles diffuse, they create a secondary effect. Since some of the mobile positive charge near the border are moving left and disappearing, the fixed negative charge in this region is no longer matched by mobile positive charge; similarly the fixed positive charge on the left side of the boundary is also exposed. This means around this boundary we have now have a net negative charge on the right, and a net positive charge on the left. Remembering back to the first chapter, this charge separation in the middle of our diode causes

[4] There are a lot of restrictions here about the types of material and how they are connected together. Generally both sides of the diode are the same material, a semiconductor, that is made conductive by adding a small amount of another material into it which is called doping. One side is doped to make the mobile carriers positive, and the other side is doped to make the mobile charges negative

[5] Another more complicated process that we will talk about again later in this section

an electrostatic force field to form (often called an electric field), and applies a force on the mobile charges.

Question: What is the direction of the electrostatic force? Which way does it push the mobile carriers?

There are extra negative charges on the right, and positive charges on the left of the junction right where the two materials meet. This charge separation pushes negative charges to the left, and positive charges to the right. Notice that this force is opposite the direction that the charges were going as a result of diffusion. As diffusion continues, the field gets larger until this field is large enough to stop the diffusion current caused by the concentration gradient. One way to think of this situation is to consider the total current to be the sum of the diffusion current and the current caused by the electric field, which is called the drift current. In this view the diffusion current causes the electric field to grow, which increases the drift current, until the drift current is equal and opposite to the diffusion current. At this point the net current through the device is zero.

Question: We said that charge separation causes a voltage to be produced. Since there is charge separation in a diode, can we measure the voltage it causes with our voltmeter?

The electric field in the device does create a voltage difference between the mobile positive carrier and mobile negative carrier materials. At first it seems like we should be able to measure it. Unfortunately these types of "voltage" differences happen whenever you connect different materials together. So while there is a voltage difference in the middle of the diode, there also is a voltage difference between the end of the diode material and the wire that connects to it. So if the leads of the diode are the same material, you won't be able to measure any voltage difference between the leads, even if there is a voltage difference inside the material.

Question: What happens if we apply an external voltage to the diode?

When we apply an external voltage, we either make the voltage difference inside the diode larger, which increases the internal field, or smaller, which decreases the field. When the field gets larger, no current can flow, since now the field pushes positive charges to the right and negative charge to the left. But there are no mobile positive carriers on the left to move right, or mobile negative charge on the right to move left. However, when we apply a positive voltage to the side that has mobile positive charge, we decrease the internal voltage. This decreases the drift current, so there is a net diffusion current, and current will flow through the device. Since the diffusion current is very large (there is a very large concentration gradient in a diode), as you increase the voltage the diffusion current becomes very large.

If you want to know more about how this works, you should take a class like EE 116.

2.6.2 The physics behind light (deep background)

Light is all around us. We use it every day, but when you dig into the physics behind light, things get a little strange (it is quantum stuff again). Let's start with the not strange stuff. Light is an electromagnetic wave that can travel through the air/space like radio, TV, or cell phone signals. These types of signals are characterized by their frequency, which also sets the length of a wave that

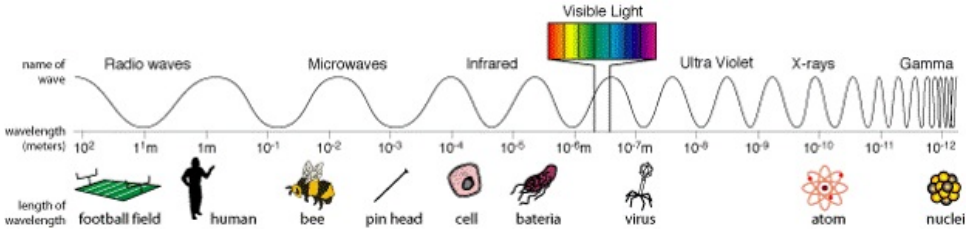


Figure 2.12: The electromagnetic spectrum. The wavelength ranges from multiple football fields in length to 10^{-12} for high energy gamma rays.

is traveling in space. Since these waves all travel at the speed of light, $3 \cdot 10^8$m/s, short wavelengths require very high frequencies. Your cell phone runs between 1-5GHz, and has a wavelength that is a few tenths of a meter. Light runs at 100s of THz, and has a wavelength that is less than 1 µm. The figure shown below shows the wavelength of different types of electromagnetic waves.

What makes light strange is that while light is a wave, it is also a particle called a photon. This duality is part of quantum mechanics, and has been validated in many experiments. We know that photons exist, since when we measure very dim light, we record the arrival of individual photons. The photon nature of light is interesting since this means that light is quantized into photons, and the energy that a photon carries is precisely set by the frequency of the light it is transmitting.

$$E = \frac{hc}{\lambda}$$

where h is Planks constant, $6.626 \cdot 10^{-34}\ J \cdot s$, and c is the speed of light, $3 \cdot 10^8$ m/s.

Since we are interested in the exchange of energy between a photon and a charge particle, either electron or its positive counterpart, we can measure this energy in electron voltage (eV). This is the energy needed to raise the potential energy of an electron by 1V. In these units, $hc = 1.24eV \cdot$ µm.

2.6.3 LEDs (deep background)

When we forward bias a diode, mobile + and - charges move across the junction between the two materials (this is the diffusion current we mentioned in the previous section), and then combine with a charge of the opposite type. Since there was an internal voltage built up across this junction, when the positive carrier makes it across the junction, it has more energy than the mobile negative charges. It generally loses this energy by converting it to heat, but in some materials it loses this energy by emitting a photon. These materials are used to make LEDs. When this happens the color of the photon is related to the energy that the particle loses, which is related to the internal voltage in the diode, which is related to the forward voltage of the diode.

Question: What voltage drop do we need to release a photon of visible light?

Visible light is roughly in the range of 400 to 750 nm, with red at about 700 nm. Using the value of hc = 1.24 eV · µm, red light would need a energy of $\frac{1.24eV}{.7} = 1.8eV$, and blue light would need an energy of $\frac{1.24eV}{.4} = 3.1eV$. Thus, we should expect the on-voltage of blue LEDs should be around 3V, and red LEDs around 2V.

Question: How do we get white light out of an LED?

This requires a trick, since a single LED can only output one color of light, and as we showed at the first lecture, white light is actually a collection of many different colored light. To solve this problem, white light LEDs use the same trick that was used in fluorescent lights — they use material called a phosphor that can absorb blue or ultra violet light, and then emits lower energy (green, yellow and red) photons. The quality of the phosphor has improved over time, so modern white LED have a color spectrum that is similar to sunlight (or incandescent lights if that is what you want). If you look at white LEDs, you'll usually see some kind of yellow spot where the diode itself is. Since a white LED is really a blue LED in disguise, the forward voltage for a white LED is also about 3V.

Question: How should LEDs be wired to ensure that they have the same brightness?

When thinking about controlling LED brightness, it is important to remember two things. First, the amount of light an LED puts out is proportional to its current. A photon is created by current flowing across the junction, so if you double the current you double the brightness of an LED. Thus, to control an LED's brightness, you need to control its current. Second, it is important to remember that LEDs are diodes, which means that their current-voltage relationship is diode-like, which means it has no current until you reach the forward voltage, and then it can support large changes in current with almost no voltage change. That means you can't really control the brightness of an LED with just a power supply.

Ideally you would like to feed an LED with a current source, since the value of that source would directly control the brightness. Some circuits do exactly that. But if you have a voltage source or battery and not a current source, you can also solve the problem by putting a resistor in series with the LED, and connecting that series combination to the battery. Since the battery has a defined voltage, and the LED has a defined voltage, the voltage across the resistor will be the battery voltage minus the forward voltage of the diode. The current through the resistor will be this difference/R, and since the diode is in series with the resistor, that same current flows through the diode.

We've seen that current flowing through an LED releases photons, so we might ask: is it possible to generate current by shining light on an LED? In other words, if an electron can fall across a bandgap and release a photon, can an incoming photon push an electron up across the bandgap?

The crazy answer is yes, this does in fact work. In fact, this is the principle that makes solar panels work.

2.6.4 Solar Cells - Capturing Light

A solar cell is a diode that is made generally from silicon, which has a forward drop of around 0.7 V. If a photon with enough energy is absorbed into the solar cell, it can create a mobile + and - charge pair. Normally if this happens on the side with many mobile - charges, the + charge will just combine with another - charge and nothing happens. But if the light is absorbed close to the junction between the two materials, there will be an electric field, and the two carriers will be pushed in different directions, and the mobile + charge will move to the side with more mobile + charges, and the mobile - charge will be pushed to the side that has mobile - charges. This means that each photon that strikes the solar cell near the junction generates a current between the

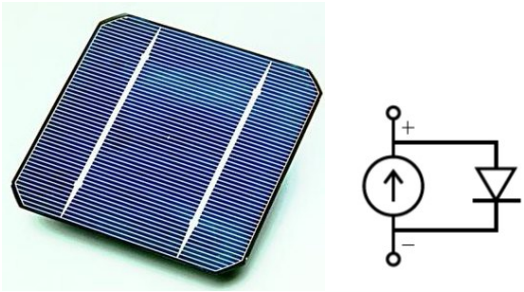


Figure 2.13: A picture of a solar cell on the left, and the circuit model of the solar cell on the right. When light hits the solar cell it generates a current which is represented by the current source. The value of the current is proportional to the photon flux, which is proportional to the light intensity.

positive and negative leads, and can be represented by a current source, as shown in Figure 2.13. Solar cells are made with the junction close to the surface, so that most of the light gets absorbed close to the junction and can generate current.

This means if you take a solar cell out in the light and connect your current meter between its two leads, the low voltage across your current meter (<100mV) will keep the diode voltage below its turn on voltage (which is around 0.7 V), keeping the diode off. The only current you will measure is current generated by the photon flux.

Question: What is the maximum voltage a solar cell can produce?

If you put a solar cell in the light, photons will turn on the current source. Since we know that charge is conserved, if current is flowing through the current source, and there is nothing connected to the solar cell terminals, this current needs to flow somewhere else. Remember a current source will create any voltage needed to maintain its current, so it will increase its voltage until the net current flow is zero. This balance position is reached when the diode in parallel to the current source turns on. When the voltage across the diode is equal to the forward voltage, all the photon-generated current will flow back though the diode. If you measure the solar cell in this position, you will measure the forward voltage of the diode across its terminals: the current source turned on the diode.

The iV characteristics of this device can be found by adding together the current from the current source and the current from the diode and is shown in Figure 2.14. The only tricky part is that the current source flows out of the positive terminal and into the negative terminal, which means from the conventional labeling for the diode this is a negative current. This added current means that this device can operate in three different quadrants, and one of them can supply power to the circuit. In the lower right quadrant, the solar cell is providing power to the circuit.

This iV plot makes it very simple to see the current we will measure when we short out the solar cell with our current meter (the short circuit current), and what voltage will appear if we don't allow any current to flow out of the device (the open circuit voltage). The open circuit voltage is the point where the net current from the device is zero, and the short circuit current is the point on the curve where the voltage across the device is zero. While these points are easy to measure, neither of those points allow you to extract much power from the solar cell. Since power is iV, the power you extract in both of these situations is about zero. One has roughly zero voltage, and one

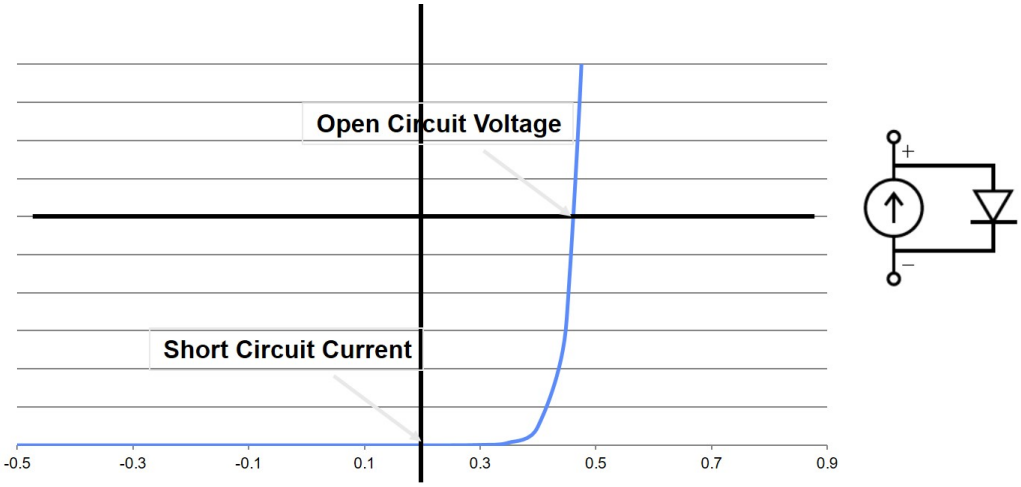


Figure 2.14: The iV curve of a solar cell when light is shining on it. Notice that this is just the iV curve for a diode shifted down by the value of the current source: we just added a constant current to the diode current.

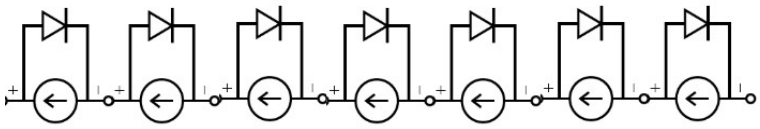


Figure 2.15: How a solar cell can generate a higher voltage by stacking many cells in series.

has zero current.

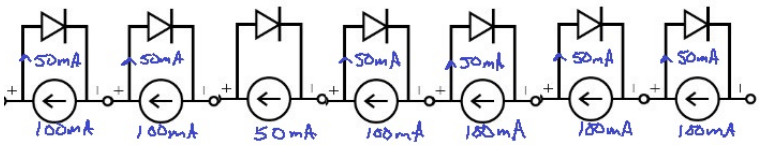
If we use our idealized model for the silicon diode, that is the current through it is zero until the forward voltage is V_F, then the open circuit voltage would be V_F. If I moved just a hair below that voltage, I would still be able to extract all of the available current. In this case the max power I could get from the solar cell is $I_{SC} \cdot V_{OC}$. While this is a guaranteed not-to-exceed power number, the maximum power you can get from a solar cell is only slightly smaller than this number.

While a large solar cell can generate a lot of current, the max voltage is limited by the forward voltage of the diode, which for silicon is around 0.6V. Since our battery is 3.7 - 4V, it will never generate enough voltage to charge the battery. There are a couple of solutions for this problem. Most solar cells increase the output voltage by putting a number of solar cells in series, as shown in Figure 2.15. This approach works and is how most solar panels are made, but it still has a major weakness: the only path for the photo-current is through the series connected current sources.

Question: What happens if 6 of the current sources in Figure 2.15 generate 100 mA but one of the cell is partially blocked and only generates 50 mA?

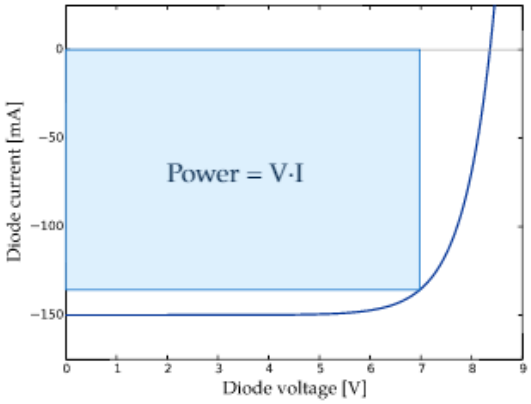
In this circuit the diodes only conduct current from left to right, while the current sources are producing current that flows from right to left. The only path for this current is through

the next current source, so if one of the current sources in the string is only 50 mA, the total current from the entire string will only be 50 mA. When the 100 mA current reaches this device, half of the current will flow through the weak current source, and the other half will flow to the right, through the diode that is in parallel with the current source as shown below.

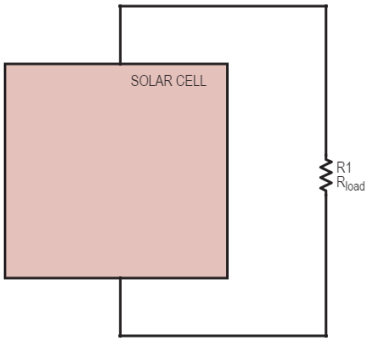


Problem 2.7

We will solve some problems related to a solar cell which has the specifications as shown in the graph below. It provides power of 1W at an output voltage of 7V.



1. How much current does the solar cell provide when its output voltage is 7V?
2. Consider the solar cell connected to a load resistor as shown below. What value of R_{load} will draw the maximum power from the solar cell?



3. Consider the same solar cell. Again assume it is connected to a load resistor, and you do not know what the value of the resistor is, but you measure the voltage across the resistor and it is 3V. What is the current flowing through the resistor? Knowing this, what must be the value of the resistor?

2.7 Solutions to practice problems

Solution 2.1:
V_1=5V, V_2=5V

Solution 2.2:
I_1=1A, I_2=1A

Solution 2.3:
I_1=50mA, I_2=22.7mA
V_1=100V, V_2=220V

Solution 2.4:
Ideal diodes
D1 and D4 are on, while D2 and D3 are off. It would be a helpful exercise to check the current through the diode for the case of it being on, and then being off, to convince yourself that for the diode to be on, the current through it must be positive.

Non-ideal diodes
Solution 2.5:
Only D3 is turned on. The current through I_1 is simply $I_1 = \frac{V}{R} = \frac{5-3}{100} = 20mA$.

Solution 2.6:
Non-ideal diodes

For circuit A:
Diodes D1 and D2 are both turned on.
$I_1 = \frac{5V-1V}{50\Omega} = 80mA$
$I_2 = \frac{5V-2V}{50\Omega} = 60mA$.
The voltage source provides a fixed voltage which is always greater than their forward voltage, and provides however much current the circuit requires. Hence, both diodes are always turned on.

For circuit B:
This circuit is more complex than it initially appears. It is important to understand here that the current source can have ***any*** *voltage across it. By making different assumptions about which diodes are on, the voltage across the current source will change. Therefore, for this kind of circuit, we first make a "guess" of which diodes are on, then check our assumptions. Essentially, this is a chance for you to practice applying the steps described above to more complex diode circuits.*

To solve this particular circuit, the steps we take are as follows:
1. Assume D4 is ON and D3 is OFF. (Arbitrary choice - you can always choose another option, but preferably one easier to solve)

2. Solve the circuit with this assumption: The voltage across the current source is: $V_I = 2V + 20mA \times 150 = 2V + 3V = 5V$ However, if this is the case, then the voltage across D3 would be $> V_{f_3}$, which means that D3 must also be on.
Hence our assumption is incorrect. D3 is also on.

3. Solve the circuit with this assumption:
- Let's call the voltage across the current source V_{tot}.
- Voltage across $R_3 = V_{R3} = V_{tot} - 1V$
- Voltage across $R_4 = V_{R4} = V_{tot} - 2V$
- By Kirchoff's current law: $20mA = \frac{V_{R3}}{R3} + \frac{V_{R4}}{R4} = \frac{V_{tot}-1V}{R3} + \frac{V_{tot}-2V}{R4}$
- Solving the above equation, we find that $V_{tot} = 3V$. This is enough voltage to turn both the diodes on so this assumption is correct.

4. Thus:
$I_3 = \frac{3-1}{150} = 13.3mA$
$I_4 = \frac{3-2}{150} = 6.6mA$

Solution 2.7:
1.) Output power is 1W at 7V, so:

$$I_{out} = \frac{P}{V} = \frac{1}{7} = 142mA$$

2.) Maximum power is drawn when $V = 7V$ and $I = 142mA$. The only resistor value for which that is true is $R = \frac{V}{I} = \frac{7}{142mA} = 49.3\Omega$.

3.) You can simply read of the specifications graph. At an output voltage of 3V, the solar cell delivers 150mA of current. Hence the current flowing through the resistor is 150mA. The resistor value must then be 20Ω.

Chapter 3

Solving for Voltage and Current

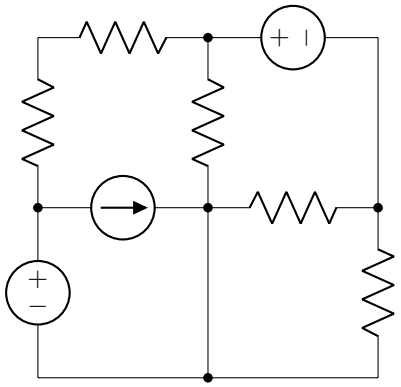


Figure 3.1: Complex circuit we will use as an example to show how to use nodal analysis to solve for device voltages and currents.

We have all the tools we need to find the voltage across, and the current through, every device in the circuit. Chapter 1 provided the constraints on the voltages (energy conservation) and currents (charge conservation), also known as Kirchoff's current and voltage laws that every circuit must follow. Chapter 2 provided the relationship between current and voltage for a number of different types of electrical devices. This chapter provides the recipe for how to use this knowledge to generate these device voltages and currents.

Consider a circuit consisting of resistors, current sources and voltage sources like the circuit in Figure 3.1. This circuit is complex and creates many simultaneous equations: one for each voltage loop (KCL), one for each current junction (KVL), and one for each resistor (Ohm's law). But solving a circuit like this *can* be done using a procedure known as *nodal analysis*.

Nodal analysis always works, and the recipe that you need to follow is not very complex. It will be the default way you analyze circuits. The downside of this method is that it generates a set of linear equations that need to be solved to find the desired answer. While this is easy for a computer to do (and it is the basis for almost all computer simulation or circuits), it gets annoying for humans when the number of variables get large. For this reason, after describing nodal analysis,

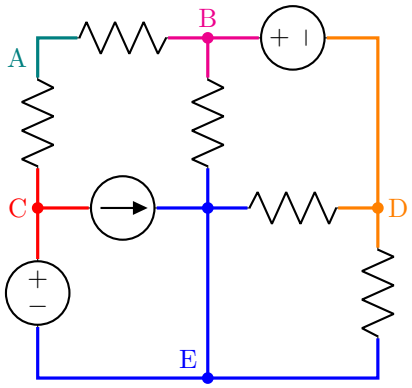


Figure 3.2: The same circuit shown in Figure 3.1 where each node of the circuit is shown in a different color. Notice that they are only 5 nodes (A-E), and that their is a voltage source in parallel with a current source (between nodes A and C), and two resistors in parallel (between nodes D and E), and two resistors in series (node A has only the two resistors attached to it)

we will use some of our insights about circuits to simplify how we analyze circuits. These insights also can be used to help reason about how device changes will affect the overall circuit performance.

3.1 Nodes and Node Voltages

As we described in Section 1.3.1, a **node** is a place where two or more devices meet. Recall that in our circuit model abstraction all device terminals that connect to the node have the same potential. Said differently the voltage difference between any two terminals connected to the same node is zero. (the lines connecting devices have zero resistance, there is zero voltage drop along the lines that make up a node). Thus even if a node has many line segments, it is a single node with a single potential.

As an example, return to the circuit in Figure 3.1. A good way to find nodes is to pick a point on a line, and follow the line in all directions possible until you run into another component. One node might only connect two components, or it might connect many. Figure 3.2 shows the same circuit, with each of the five nodes highlighted in a different color. Once you have identified all the nodes, it is easier to see which devices are in parallel (they share two nodes), and which are in series (they exclusively share one node).

Recall from Chapter 1 that voltage is a measure of potential difference *between the two terminals of a device*, and we call this voltage difference the device's voltage, or the voltage difference (voltage drop) across the device. From conservation of energy, we also know that the sum of device voltages around any loop will be zero. Said differently, this means one can compute the potential difference between any two nodes by summing the device voltages between those node, and that voltage won't depend on the path they take. This means we can compute the voltage difference between node D and E in Figure 3.2 by either using the voltage across the resistor that connect those two nodes, or by adding the voltage across the resistor to node B and the voltage source between B and D.

Nodal voltages

Since the voltage difference between two nodes doesn't depend on the path taken, we can start talking about the voltage difference between any two nodes. The only tricky part of computing the

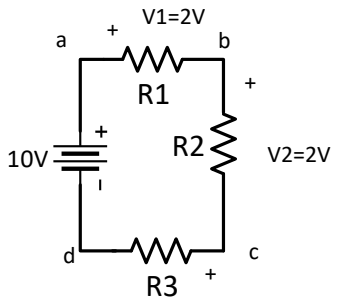


Figure 3.3: To find the voltage difference between any two nodes, you need to add the device voltages with their reference direction aligned with your path and subtract the device voltages with reference directions that are opposite to your path direction. Say you are interested in computing the voltage from node 'b' to node 'd'. A positive voltage means that node 'b' is at a higher potential than node 'd'. So the reference direction of our path has the '+' terminal pointing to node 'b'. That means the voltage across devices where the '+' reference direction of the device is closer to b than the '-' end should be added to the total, and the voltages of devices where the '-' end of the device is closer to 'b' than the '+' end should be subtracted from the total. Traveling from node 'd' to 'b' through the voltage source: add 10V of the voltage source since it is aligned, and then subtract 2V of the resistor (since the '-' terminal is closer) to yield a 8V difference. The voltage from node 'b' to node 'd' is 8V.

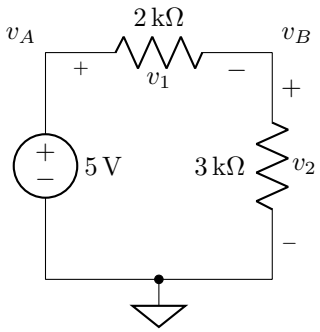


Figure 3.4: In this circuit $v1$ and $v2$ are device voltages and v_A and v_B are node voltages. One can derive the device voltages from the node voltages and vice versa.

voltage around a loop, or nodal voltages is to deal with the device reference directions correctly. Figure 3.3 shows this issue. To find the voltage from node 'b' to node 'd' add the voltages that are aligned with your path, and subtract the voltages that are aligned in the opposite direction. If you want the voltage from node 'b' to node 'd' than reference directions aligned with your path will have the '+' reference direction terminal closer to 'b' (the node we are trying to measure) than its '-' terminal. In this figure the voltage source reference directions match this path, but R1's reference direction is backward. Thus the voltage between 'b' and 'd' is $10\,\text{V} - 2\,\text{V} = 8\,\text{V}$.

It doesn't make sense to talk about the voltage "at" a node since voltage is a measure of potential difference. Nonetheless, because talking about the voltage of a node relative to another node can get cumbersome, it's common in electrical engineering to choose one node as a *reference point*, and measure the voltage of all the other nodes relative to this reference. This reference node is called *ground*, often abbreviated as GND, and is indicated on circuits with this symbol: ⏚.

Thus, when we say "*the voltage at node A is 3.4 V*", what we really mean is "*the voltage difference between node A and GND, with the reference direction defined so that + is at node A, is 3.4 V*". Since nodal voltages are defined as the voltage difference between that node and GND, the voltage of the node GND is, by definition, 0 V: the voltage difference between a node and itself is always zero.

Node voltage vs device voltage

A voltage "at" a node is called a **node voltage**. This is distinct from the voltage across a device, which is called a **device voltage**. Please don't confuse the two. You find the device voltage by subtracting the node voltage connected to the device terminal labeled '-' from the node voltage connected to the '+' labeled device terminal. Please don't use a node voltage as a device voltage. Unless that device is between the node and GND you will get the wrong answer. For example, in the circuit in Figure 3.4, v_1 and v_2 are device voltages, while v_A and v_B are node voltages. Following our rules for device voltages, $v_1 = v_A - v_B$, and $v_2 = v_B - 0$ (the 0 being ground).

Choosing a ground node

We can choose any node we want to serve as ground. If you move ground to a different node, all of the node voltages will change (since they're measured relative to the new ground), but the voltage differences between nodes (including all device voltages) will not change. Figure 3.5 shows the same circuit as in Figure 3.4, but with different nodes chosen as ground. In each case, the node voltages v_A, v_B, v_C are different, but the voltages across each device always remain the same. Furthermore, since the current through a device depends on the voltage *across* that device (devices don't know where ground is!), the choice of ground doesn't affect any calculations of current flowing in the

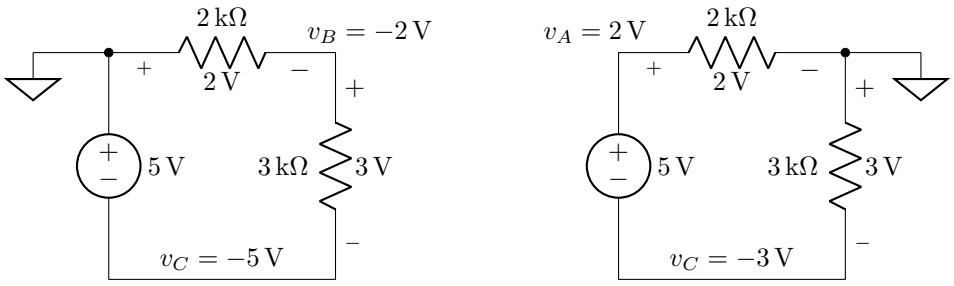


Figure 3.5: The same circuit with different nodes chosen as GND. While the node voltages change in the two circuits (since the voltage is measured against a different reference, the device voltages (and hence device currents) don't depend on the choice of reference voltage.

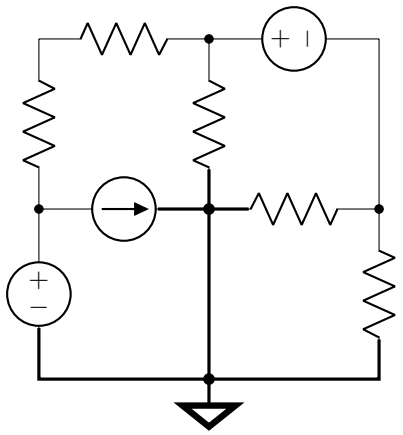


Figure 3.6: It is usually a good idea to choose the node with the most devices connected to it as the GND (reference) node, which was done in this circuit.

circuit. This makes sense: the operation of a circuit shouldn't depend on an arbitrary reference point!

It's common (but not universal) to choose the lowest-voltage node in the schematic to serve as ground. Engineers also often draw circuits so that nodes with lower voltages are at the bottom of their diagram, which often makes ground at the bottom. Another strategy is to make ground the node with the most connections to other devices, since, as we will see next, this will simplify nodal analysis. For example, in our circuit from Figure 3.2, we might choose node E, since that node connects five components (Figure 3.6).

Redrawing with ground at bottom

However, if we wanted to choose another node, we could do that too. We might instead choose node B, as shown on the left of Figure 3.7. It often helps to redraw the circuit with ground at the bottom, as illustrated on the right of Figure 3.7. In general, you can choose any node you want to be your ground node. Just make sure you mark your ground node clearly whenever you're solving a circuit.

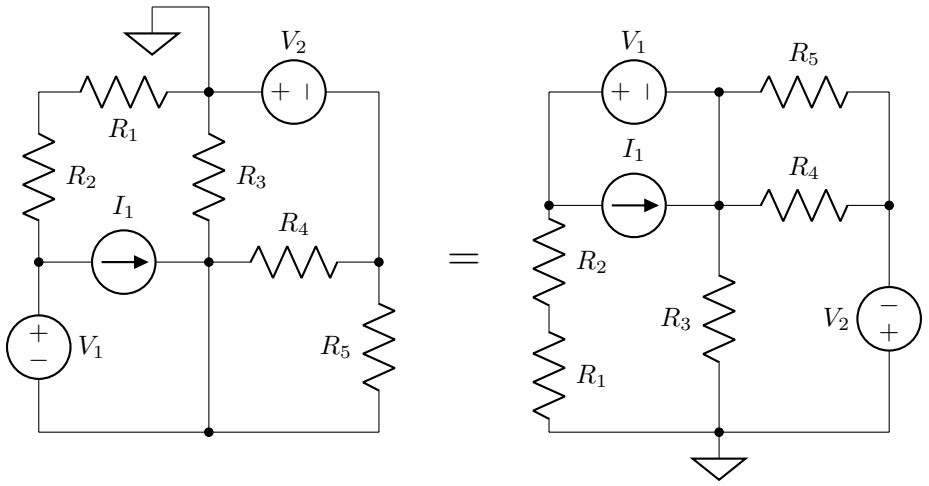


Figure 3.7: Redrawing circuit with ground at bottom. These two circuits are the same. The right circuit has simply rotated the circuit on the left to move the ground node to the bottom. If you label all the nodes, you will see that each node connects to the same devices in both circuits.

3.2 Nodal Analysis

Nodal analysis procedure

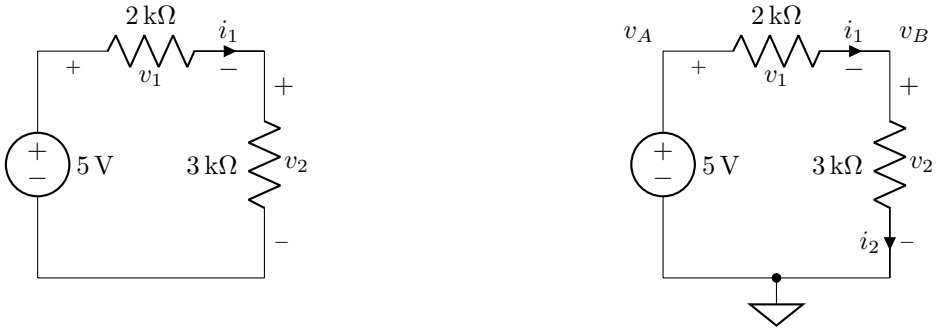
Nodal analysis is a method of solving for all voltages and currents in a circuit, which works as follows:

1. Choose a ground node, and label every unknown node voltage.
2. Label the current and the voltage reference directions for each device (and make sure the current reference direction flows into the device terminal you labeled '+').
3. For each voltage source, write an expression relating the node voltages on either side of it.
4. Write KCL equations for every node where the node voltage is not known (current that flows into a node, must flow out of it).
5. Now use the device current voltage relationships to express device current as a function of nodal voltages.
6. Solve the resulting system of simultaneous equations.

In essence, this is a systematized application of Ohm's law and Kirchoff's current law, using node voltages (which implicitly use Kirchoff's voltage law). It's not as complicated as it first looks. The key step is step 4. If you remember that it's based around charge conservation (KCL), you'll find the rest follows naturally, and before you know it you won't even be thinking of any step other than step 4. The catch is that it can be laborious, particularly the last part, where you might be solving a large system of simultaneous equations, often using techniques from linear algebra.

Nodal analysis is best demonstrated by example.

Example 3.1 Find the current i_1, and the voltages v_1 and v_2, in the circuit below left.



Solution. *Step 1: Choose a ground node and label nodes with unknown voltages.* It seems natural to choose the bottom node, as shown above right, which leaves two other nodes. We will label the voltages at these two nodes we label v_A and v_B. We don't need to label ground, because its voltage is 0 V, by definition.

Step 2: Label the current of every branch, and the voltage reference directions. Most of the device currents were already given to us, and the device reference directions are given. All we do is add a label i_2 for the current through the second resistor to the already labeled i_1, It's not generally helpful to label the current through voltage sources (since the current through them is not constrained).

Step 3: For each voltage source, write an expression relating the node voltages on either side of it. There's one voltage source, which connects to ground on one side and v_A on the other:

$$v_A - 0 = 5\,\text{V} \qquad \Longrightarrow \qquad v_A = 5\,\text{V}.$$

Step 4: For each node with an unknown voltage, use Kirchoff's current law to relate the currents entering that node. Since we know the voltage of node 'A', v_A, we can skip it (it is not unknown), so we only need to write the equation for node 'B':

$$i_1 = i_2 \qquad \textit{or} \qquad i_1 - i_2 = 0$$

Step 5: Now use the device current voltage relationships to express device current as a function of nodal voltages. We only have two device currents that we are interested in, and we need to express these in terms of the node voltages, v_A, v_B, not device voltages v_1, v_2. If step 3 fixed some node voltages, we can use those values in these equations:

$$\text{For the } 2\,\text{k}\Omega \text{ resistor:} \qquad i_1 = \frac{v_A - v_B}{2\,\text{k}\Omega} = \frac{5\,\text{V} - v_B}{2\,\text{k}\Omega}.$$

$$\text{For the } 3\,\text{k}\Omega \text{ resistor:} \qquad i_2 = \frac{v_B - 0}{3\,\text{k}\Omega} = \frac{v_B}{3\,\text{k}\Omega}.$$

Step 6: Solve the resulting system of simultaneous equations. It's normally easiest to start

with the KCL equations from step 4, then substitute the relevant expressions from step 5:

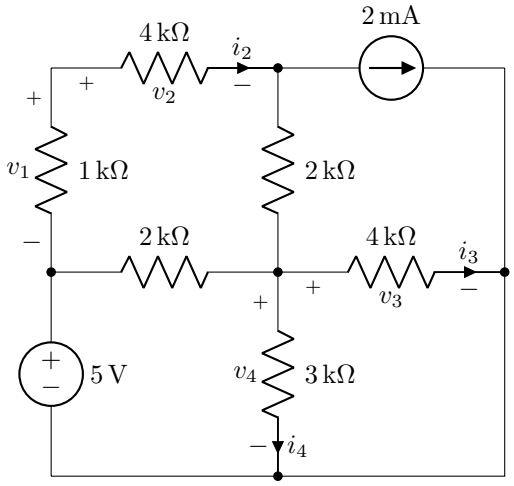
$$
\begin{aligned}
i_1 &= i_2 \\
\frac{5\,\mathrm{V} - v_B}{2\,\mathrm{k\Omega}} &= \frac{v_B}{3\,\mathrm{k\Omega}} && \text{from step 5} \\
\frac{5\,\mathrm{V}}{2\,\mathrm{k\Omega}} &= \frac{v_B}{3\,\mathrm{k\Omega}} + \frac{v_B}{2\,\mathrm{k\Omega}} \\
\frac{5}{2} &= v_B \left(\frac{1}{3} + \frac{1}{2}\right) \\
v_B &= \frac{5}{2} \times \frac{6}{5} = 3\,\mathrm{V}.
\end{aligned}
$$

It then follows that

$$
\begin{aligned}
i_1 = i_2 &= \frac{3\,\mathrm{V}}{3\,\mathrm{k\Omega}} = 1\,\mathrm{mA}, \\
v_1 &= v_A - v_B = 5\,\mathrm{V} - 3\,\mathrm{V} = 2\,\mathrm{V}, \\
v_2 &= v_B - 0\,\mathrm{V} = 3\,\mathrm{V}.
\end{aligned}
$$

In the above example, it wouldn't have been too hard just to write the same equations by inspection, without the overhead of a procedure. In the following example, the nodal analysis procedure will help keep us more organized.

Example 3.2 Find the currents i_1, i_2, i_3, i_4 and voltages v_1, v_2, v_3, v_4 in the circuit below.



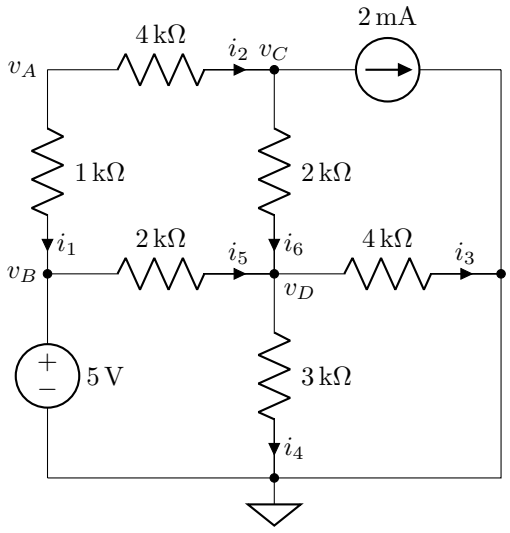
This is a circuit that you only see in textbooks or problem sets. It is much more complex (with more nodes) that any normal circuit you will need to solve. It is here just to show you that even when circuits are complicated, the simple procedure still works.

Solution.

Step 1: Choose a ground node and label nodes with unknown voltages. Again, it seems natural to choose the bottom node as the reference node, which leaves us with four nodes with unknown voltages (one we will figure out fairly quickly). We will label the nodes A-D as shown in the figure below.

Step 2: Label the current of every branch, and the voltage reference directions. We add labels for the remaining currents. To reduce clutter, we've removed the device voltage labels, which removed the reference directions in the diagram, but note that they remain unchanged. That means that:

$$v_1 = v_A - v_B, \qquad v_2 = v_A - v_C, \qquad v_3 = v_D - 0, \qquad v_4 = v_D - 0.$$



Step 3: For each voltage source, write an expression relating the node voltages on either side of it. There's one voltage source, which connects to ground on one side and v_B on the other:

$$v_B - 0 = 5\,\text{V} \qquad \Longrightarrow \qquad v_B = 5\,\text{V}.$$

Step 4: For each node with an unknown voltage, use Kirchoff's current law to relate the currents entering that node.

For node A:	$0 = i_1 + i_2$
For node C:	$i_2 = i_6 + 2\,\text{mA}$
For node D:	$i_5 + i_6 = i_3 + i_4.$

Step 5: Now use the device current voltage relationships to express device current as a function of nodal voltages. Remember to use the node voltages to generate the needed device voltages

$$i_1 = \frac{v_A - v_B}{1\,\text{k}\Omega} = \frac{v_A - 5\,\text{V}}{1\,\text{k}\Omega}, \qquad i_2 = \frac{v_A - v_C}{4\,\text{k}\Omega}, \qquad i_3 = \frac{v_D - 0}{4\,\text{k}\Omega},$$
$$i_4 = \frac{v_D - 0}{3\,\text{k}\Omega}, \qquad i_5 = \frac{v_B - v_D}{2\,\text{k}\Omega} = \frac{5\,\text{V} - v_D}{2\,\text{k}\Omega}, \qquad i_6 = \frac{v_C - v_D}{2\,\text{k}\Omega}.$$

Step 6: Solve the resulting system of simultaneous equations. Again, start with the KCL

equations from step 4, then substitute relevant expressions from step 5.

$$\text{For node A:} \qquad 0 = \frac{v_A - 5\,\text{V}}{1\,\text{k}\Omega} + \frac{v_A - v_C}{4\,\text{k}\Omega}$$

$$\text{For node C:} \qquad \frac{v_A - v_C}{4\,\text{k}\Omega} = \frac{v_C - v_D}{2\,\text{k}\Omega} + 2\,\text{mA}$$

$$\text{For node D:} \qquad \frac{5\,\text{V} - v_D}{2\,\text{k}\Omega} + \frac{v_C - v_D}{2\,\text{k}\Omega} = \frac{v_D}{4\,\text{k}\Omega} + \frac{v_D}{3\,\text{k}\Omega}.$$

Now we just need to solve this these three equations. For node A and node C equations we multiply both sides of the equation by $4\,\text{k}\Omega$. For node D equation, we need to multiply both sides by $12\,\text{k}\Omega$. This will remove all the denominator terms and leave:

$$\text{For node A:} \qquad 0 = 4v_A - 20\,\text{V} + v_A - v_C \Longrightarrow v_A = 4\,\text{V} + \frac{v_C}{5}$$

$$\text{For node C:} \qquad v_A - v_C = 2v_C - 2v_D + 8\,\text{V} \Longrightarrow v_D = \frac{3v_C - v_A}{2} + 4\,\text{V}$$

$$\text{For node D:} \qquad 30\,\text{V} - 6v_D + 6v_C - 6v_D = 3v_D + 4v_D \Longrightarrow v_D = \frac{30\,\text{V}}{19} + \frac{6}{19}v_C$$

We are almost done. Now we use the equation for v_A from node A to get ride of v_A in the second equation, and rewrite the third equation to convert the fractions to decimal numbers gives: . Doing this substitution gives

$$v_D = \frac{3v_C - (4\,\text{V} + \frac{v_C}{5})}{2} + 4\,\text{V} = 1.4v_C + 2\,\text{V}$$

$$v_D = 1.58\,\text{V} + 0.316v_C$$

Setting these two equations equal to each other provides the value v_C which than can be used in the equation for node A to find v_A and in the eqation for node D to find v_D. This gives:

$$v_A = 3.92\,\text{V}, \qquad v_C = -388\,\text{mV}, \qquad v_D = 1.46\,\text{V}.$$

Finally, to find the desired voltages and currents:

$$v_1 = v_A - v_B = -1.08\,\text{V}, \qquad v_2 = v_A - v_C = 4.31\,\text{V}, \qquad v_3 = v_4 = v_D = 1.46\,\text{V},$$

$$i_1 = \frac{v_1}{1\,\text{k}\Omega} = -1.08\,\text{mA}, \qquad i_2 = \frac{v_2}{4\,\text{k}\Omega} = 1.08\,\text{mA},$$
$$i_3 = \frac{v_3}{4\,\text{k}\Omega} = 364\,\mu\text{A}, \qquad i_4 = \frac{v_4}{3\,\text{k}\Omega} = 485\,\mu\text{A}.$$

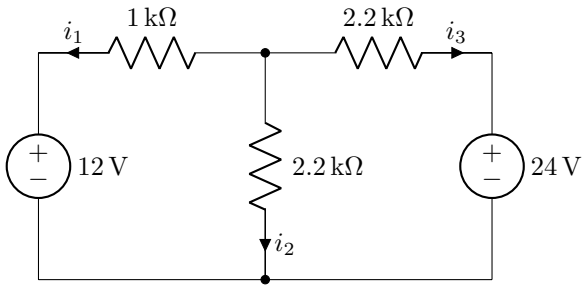
Notice that v_1 and i_1 were negative. This isn't a problem—it just means that the direction we assumed when labeling a reference direction was the opposite to the direction that positive values would match. So conventional, positive current actually flows *upwards* through the $1\,\text{k}\Omega$ resistor—which makes sense, because then i_2 is positive. It doesn't matter which way you choose your reference directions, so long as you follow the passive sign convention (current going into the

positive terminal) and you're careful to stay consistent for your entire analysis.

You might also have noticed that, around step 6, things got pretty unwieldy. Solving that system of three simultaneous equations wasn't much fun—and because you have one KCL equation per node, the bigger the circuit, the more tedious it gets. Fortunately, most of the circuits you create in making stuff stuff don't have that many nodes with unknown voltages in them, and are easier to solve than this example problem. Good makers, (efficient electrical engineers) also can use some of their insights about circuits to reduce the complexity of the circuits they need to analyze, sometimes making the circuit so simple (it only has on unknown nodal voltage) that they can solve the problem in their head. Even if the problem can't be reduced that much, it often can be simplified so the nodal analysis is not that complex. In the remainder of this chapter, we cover how to use your circuit insights (even if you don't think you have any). And remember, if you forget the insights, you can always do nodal analysis. It might take you a little longer, but it converts any circuit problem into just linear algebra.

The problem below is more like the circuit examples you will see in class. It is a circuit with 3 resistors, two supplies, but only one node with an unknown voltage. Please see if you can find the missing nodal voltage, and from that find all the currents in this circuit.

Problem 3.1 Find the currents i_1, i_2 and i_3.



3.3 Series and Parallel Resistors

We have already learned about series and parallel devices, and know these connections have special properties. We can use these special properties to reduce the complexity of the circuits we need to evaluate with nodal analysis, if those circuits have series or parallel resistors.

3.3.1 Resistors in Series

Recall from Section 1.6 that two devices are in **series** when a terminal of one is connected *only* to a terminal of the other. Because these two devices are the only devices connected at this node, the current through each must be the same. This follows from KCL: the sum of currents at the node between the two device must be equal to zero, so the same current must flow out of one device and into the other. Figure 3.8 shows a situation where two devices are in series, and a situation where none of the three devices are in series.

When two resistors are connected in series, the voltage across the series combination is just going to be the sum of the voltage across each resistor. But this voltage is $i \cdot R_1$ and $i \cdot R_2$, and since the resistors are in series, the i value is the same. This means one can factor out the i from

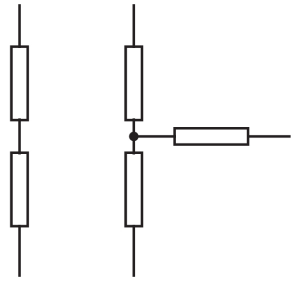


Figure 3.8: The circuit on the left has two devices in series. None of the three devices is in series for the circuit on the right. In the right circuit there isn't a node that connect only two devices together.

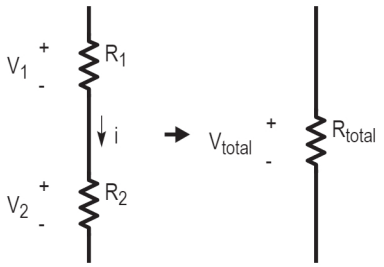


Figure 3.9: Combining resistors in series

the sum, so the voltage across the series combination of resistors is $i \cdot (R_1 + R_2)$. So the sum of the voltage across these two devices will be the same as the voltage across a single resistor of (R_1 + R_2). This is shown in Figure 3.9

$$V_1 = i * R_1$$

$$V_2 = i * R_2$$

$$V_{total} = V_1 + V_2 = iR_1 + iR_2 = i(R_1 + R_2) = iR_{total}$$

$$R_{total} = R_1 + R_2$$

The big advantage of doing this simplification is that you remove a node from the circuit, which removes one of the variable you need to solve for and one equation from your nodal analysis. While makes the linear algebra easier, its real advantage is when it makes the algebra so simple that you can solve it without really doing any linear algebra (which is true for most of the problems you need to solve).

If you have N resistors in series, $R_1..R_N$, you can repeatedly apply this method to combine a resistor to the previous equivalent resistors, so this equation generalizes to:

$$R_{total} = \sum_{n=1}^{N} R_n$$

Notice from this equation, that the total resistance along a path always increases as you add another resistor in series with it. Checking that this is true is always a good way to check your work.

3.3.2 Parallel Connections

Again as pointed out in Section 1.6 a parallel connection refers to one in which the two nodes that the first component connects to are the same two nodes that the second component connections to. Other devices can also connect to either node. The special property of parallel components is that they must have the same voltage across them. This is obvious from the fact that the device voltage is the difference in the nodal voltages. Since the two devices connect to the same two nodes, their device voltages must be the same. Some examples are shown in Figure 3.10

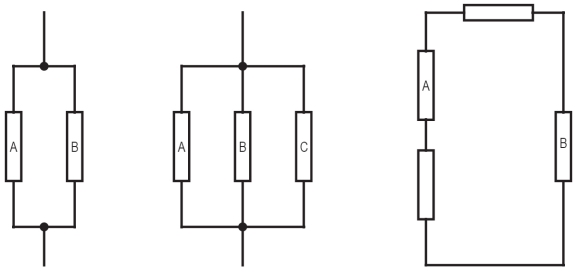


Figure 3.10: Left: A and B are in parallel; Center: A, B and C are in parallel; Right: A and B are not in parallel

When you put two devices in parallel, the current through the combination is just the sum of the currents through the two devices. Since each resistor's current is proportional to voltage, the sum of the currents will also be proportional to voltage. This linear relationship between current and voltage means that two parallel resistors look like another resistor as shown in Figure 3.11. We can find the equivalent resistance of parallel resistors is equal to $\frac{1}{\frac{1}{R_1}+\frac{1}{R_2}}$ as found by solving Ohm's Law across the two parallel components:

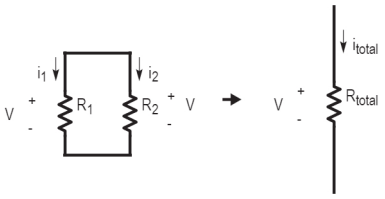


Figure 3.11: Combining resistors in series

$$V_{total} = i_1 R_1 = i_2 R_2$$

$$i_1 = \frac{V_{total}}{R_1}; i_2 = \frac{V_{total}}{R_2}$$

$$i_{total} = i_1 + i_2 = V_{total}(\frac{1}{R_1} + \frac{1}{R_2})$$

$$R_{total} = \frac{V_{total}}{i_{total}} = (\frac{1}{R_1} + \frac{1}{R_2})^{-1} = \frac{R_1 \cdot R_2}{R_1 + R_2}$$

This equation can be generalized for multiple resistors in parallel:

$$R_{total} = (\sum_{n=1}^{N} \frac{1}{R_n})^{-1}$$

From this equation, the equivalent resistance of resistors in parallel will be smaller than any of the individual resistors. With parallel connections, the current has more possible paths to take, which means that the ability of the circuit to conduct (conductance) increases, and resistance, which is the reciprocal of conductance, decreases.

3.3.3 Combining Series and Parallel Resistors

Let's try out what we know about series and parallel resistors to solve for the equivalent resistance between A and B in the circuit shown in Figure 3.12.

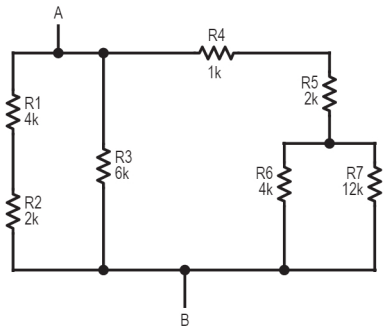


Figure 3.12: Find the equivalent resistance between points A and B

Let's start by identifying what we know are resistors in series and parallel. R_1 and R_2 are clearly in series, since one terminal of R_1 is connected directly to a terminal of R_2, with nothing else in the middle. Similarly R_4 and R_5 are in series. We know that we can combine two resistors in series into one resistor by simply adding the resistances, resulting in the circuit in Figure 3.13.

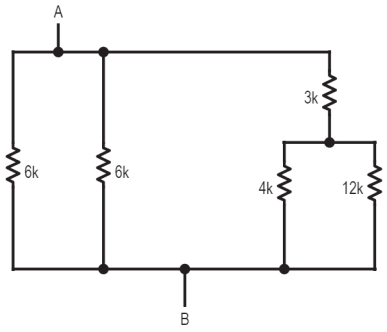


Figure 3.13: Find the equivalent resistance between points A and B - series resistances combined

Now let's try simplifying some parallel resistors. We can see that R_6 and R_7 are in parallel, since both of their terminals are connected to each other. We can use our parallel resistance equation, $R_{total} = \frac{R_6 \cdot R_7}{R_6 + R_7}$, to combine these two resistors into a single 3kΩ resistor. Similarly, R_3 is now in parallel with the combined R_1 and R_2. When two resistors of the same value are parallel to each other, the resulting equivalent resistance is just equal to half of the original resistor value (you can see this by solving the original equation: $\frac{R^2}{2R} = \frac{R}{2}$). Therefore, we can combine the left two branches of the circuit into a single 3kΩ resistor. The result is in Figure 3.14.

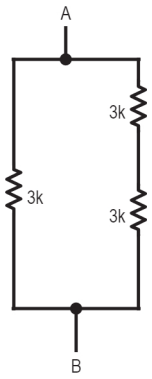
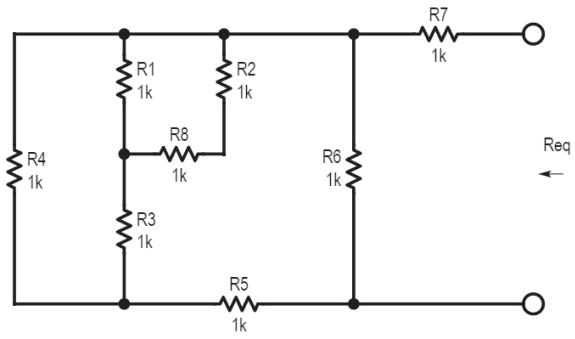


Figure 3.14: Find the equivalent resistance between points A and B - parallel resistances combined

This resulting circuit is pretty simple. The two resistors on the right are in series, so we can just add them to get a 6kΩ resistor. Then we have a 3kΩ in parallel with a 6kΩ, so we can use our parallel resistance equation to get an overall resistance of 2kΩ. We've now simplified the circuit into what's shown in Figure 3.15.

Figure 3.15: Final combined resistance

Problem 3.2
Find the equivalent resistance of this entire circuit between the two indicated nodes. (Find R_{eq}.)



3.4 Voltage and Current Dividers

Sometimes, we have two resistors in series, but we need to find the voltage at the node in the middle of the circuit, or we have two resistors in parallel but we need to find the current through one of the resistors. When we combine these resistors, the node/current of interest disappears. There is no need to panic, since we can use the information we found after the reduction to eventually find the desired parameter. For example we can do the series reduction, find the current through the devices, and then use this current back in the initial circuit to find the desired missing voltage. After we use this technique a couple of times, we see a pattern in the resulting answer. These circuit produce an output voltage or current that is a fraction of the input voltage or current. This result is a useful function, which is one of the reasons that they occur in many circuits. As a result these circuits are called voltage and current dividers.

3.4.1 Voltage Dividers

Figure 3.16 shows a simple voltage divider which is made of two resistors connected in series. We want to solve for the voltage drop across each resistor. Let's first solve this by nodal analysis, since we are sure that this approach will work. In this problem there is only one unknown nodal voltage, V_{mid}. Writing KCL for this node (assuming the device currents both flow down):

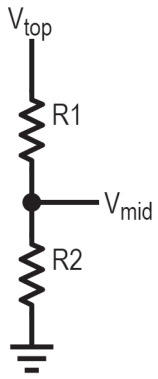


Figure 3.16: A voltage divider

$$\frac{V_{top}}{R_1} - \frac{V_{mid}}{R_1} = \frac{V_{mid}}{R_2}$$

$$V_{mid}(\frac{1}{R_1} + \frac{1}{R_2}) = \frac{V_{top}}{R_1}$$

$$V_{mid} = V_{top}\frac{\frac{1}{R_1}}{\frac{1}{R_1} + \frac{1}{R_2}} = V_{top}\frac{R_2}{R_1 + R_2}$$

Notice that the output voltage is a fraction of the input voltage, and that fraction is the value of the bottom resistor, over the total resistance of the chain. This makes sense, and we could have generated this result directly. The current flowing through the combination is the supply voltage

divided by the total resistance, $R_1 + R_2$. The output voltge is just this current times R_2. Hence the output voltage should be $V_{top}\frac{R_2}{R_1+R_2}$. If most of the series resistance is from the bottom resistor, it will have most of the voltage across it. But as this resistance gets smaller (compared to the total), the voltage across the resistor gets smaller too.

If we want to solve for the voltage across R_1 instead, we would need to find $V_{top} - V_{mid}$. We can then rewrite the equations as follows:

$$V_{mid} = V_{top}\frac{R_2}{R_1 + R_2}$$

$$V_{top} - V_{mid} = V_{top} - V_{top}\frac{R_2}{R_1 + R_2} = V_{top}(1 - \frac{R_2}{R_1 + R_2})$$

$$V_{top} - V_{mid} = V_{top}\frac{R_1 + R_2 - R_2}{R_1 + R_2}$$

$$V_{top} - V_{mid} = V_{top}\frac{R_1}{R_1 + R_2}$$

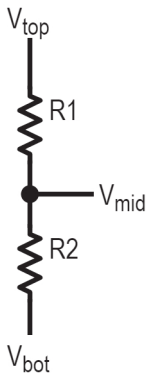


Figure 3.17: Another voltage divider

Sometimes you will have a voltage divider situation but the bottom voltage won't be ground. Instead, let's assume it is V_{bot}. While this situation seems more complicated than the grounded case, it is no harder to solve. You just need to remember that all voltages are relative. We know the voltage drop across R_2 is going to be:

$$(V_{top} - V_{bot}) \cdot \frac{R_2}{R_1 + R_2}$$

And this is equal the voltage $V_{mid} - V_{bot}$. So V_{mid} is just going to be:

$$V_{mid} = V_{bot} + (V_{top} - V_{bot}) \cdot \frac{R_2}{R_1 + R_2}$$

Note that if $V_{bot} = 0$, this equation simplifies to the same thing we had earlier. Basically, this is telling us that we need to multiply the resistance ratio by the voltage difference across the entire

divider, then add the offset from the bottom voltage. In this case, the voltage across the top resistor is equal to:

$$V_{top} - V_{mid} = (V_{top} - V_{bot})\frac{R_1}{R_1 + R_2}$$

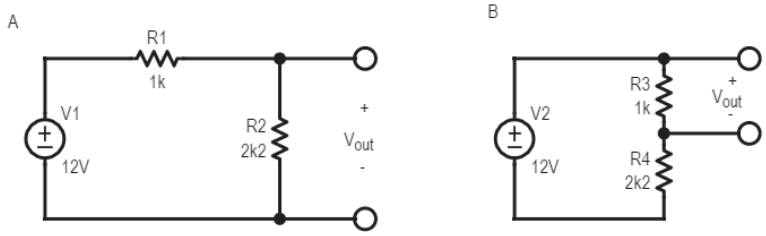
Again, this is the same equation we had earlier, but looking at the voltage difference between the resistors.

In general, the voltage across one of the components is equal to the resistance of that component divided by the sum of the two resistances. If we have multiple components connected in series between V_{top} and V_{bot}, this equation generalizes as shown below:

$$V_{R_x} = (V_{top} - V_{bot})\frac{R_x}{\sum_{n=1}^{N} R_n} + V_{bot}$$

If R_1 and R_2 are multiple components connected in series or parallel, we can find the equivalent resistance of these two sets of components and solve from there.

Problem 3.3
Find V_{out} in the following two circuits.



3.4.2 Current Dividers

In a simple current divider made up of two resistors connected in parallel, we want to solve for the current through each resistor. In a parallel circuit, the voltage drop across each component is the same. We can use this to write equations for the circuit as shown below:

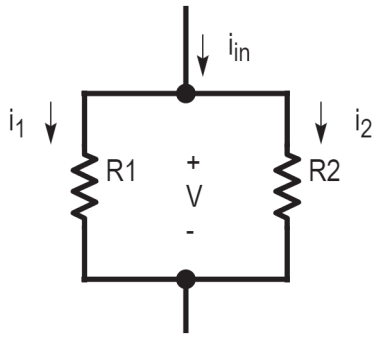


Figure 3.18: Another voltage divider

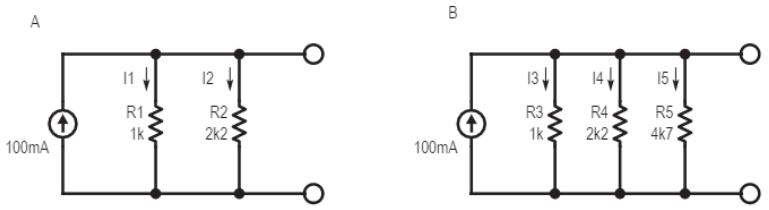
$$V_{total} = i_1 R_1 = i_2 R_2$$
$$i_2 = i_1 \frac{R_1}{R_2}$$
$$i_{total} = i_1 + i_2 = i_1 + i_1 \frac{R_1}{R_2}$$
$$i_{total} = i_1 (1 + \frac{R_1}{R_2}) = i_1 \frac{R_1 + R_2}{R_2}$$
$$i_1 = i_{total} \frac{R_2}{R_1 + R_2}$$

We can flip these equations to find the current through the opposing resistor, $i_2 = i_{total} \frac{R_1}{R_1+R_2}$. This essentially tells us that the fraction of the original current through one resistor is equal to the other resistance, divided by the sum of the two resistances. Notice that equation looks similar the the voltage divider equation, for a current divider, the numerator is the opposite resistor, not the resistor we are measuring the current through. This makes sense, since as the resistance of the resistor increases, it is harder for the current to flow, so the current should go down.

If we have multiple resistors in parallel, the fraction of the current through one resistor is equal to the parallel combination of the other resistances, divided by the sum:

$$i_{R_x} = i_{total} \frac{(\sum_{n=1}^{N} \frac{1}{R_n} - \frac{1}{R_x})^{-1}}{(\sum_{n=1}^{N} \frac{1}{R_n})^{-1}}$$

Problem 3.4
Find the currents in each of the resistors in the following circuits.



3.5 Superposition

Linear circuit have one more property that is sometimes useful for trying to understand how a circuits works. It is especially useful when you are trying to decide how a specific input (voltage or current source) affects a voltage of interest. The important property is that the voltage and currents in a circuit that only uses resistors, voltage and current sources (no diodes) are linear on the input sources. If you look through our nodal analysis procedure, or any of the results, you will always find that the voltage for any node can always be expressed as a weighted sum of each the current and voltage sources:

$$V_A = k_1 \cdot V_1 + k_2 \cdot V_2 ... + R_1 \cdot I_1 + ...$$

where V_i are the voltage sources in the circuit, I_i are the current sources in the circuit, and k_i and R_i are the proportionality constants. This means this output voltage depends linearly on each of the input sources: if you double the value of an input source, its contribution to the nodal voltage doubles. Since each contribution doesn't depend on the other sources, we can find the output voltage by computing the output from each source independently and then adding these contributions together. Basically, it means that we will get the same result by solving the entire equation at once or solving different sources separately and then adding the results together.

In most situations, solving the circuit multiple times for each source is more work than just solving the circuit once, and getting the final answer. But in some situations we are interested in figuring out how much of the output is caused by this current source, or what happens to the output voltage when this voltage source increases by 50%, and in these cases it is often easier to set the other sources to zero.

If you are going to use this method it is critical that you understand how to model a voltage source set to 0 V, and a current source set to 0 A, since they are very different. A voltage source with zero volts across it is just a wire, since the two terminals of the voltage source have the same potential. So if you set a voltage source to zero, it shorts together the two nodes it connects to. A current source with zero current in it, does the opposite. Since no current flows through the component, it becomes an open circuit, and disappears from the circuit.

We can use this analysis method to explain the nodal voltages for the circuit shown in Figure 3.19. When the current source is set to zero, the circuit becomes a simple voltage divider, and

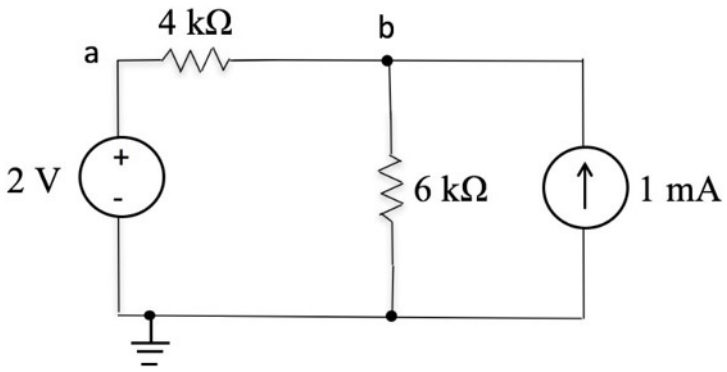


Figure 3.19: A simple circuit with two sources. While this is not hard to solve with nodal analysis, since node 'b' is the only node where the voltage is not known, it can also be solved by superposition. When the current source is set to zero, the circuit is a simple voltage divider, so v_b would be equation to 6k/10k 2 V which is 1.2 V. When the voltage source is set to zero, the two resistor become in parallel, and the parallel combination is 2.4K. 1 mA flowing through 2.4 kΩ is 2.4 V. So the total voltage is 3.6 V.

the voltage at node 'b' is 1.2V. When the voltage source is set to zero the two resistors become in parallel and the voltage at node 'b' is 2.4V. So the final voltage is 3.6V. While the nodal analysis for this circuit is not hard, using superposition made the analysis pretty simple (if you recognized the voltage divider).

There are other situations where nodal analysis might be faster. Figure 3.20 is an example. We can always use straight nodal analysis. For this circuit, we could solve for the voltage at the node at the top of the circuit - let's call it V_A. We could write our KCL equations as shown below:

$$\frac{V_A - 5\,\text{V}}{1\,\text{k}\Omega} + \frac{V_A}{4\,\text{k}\Omega} + 2\,\text{mA} = 0$$
$$5V_A - 20\,\text{V} + 8\,\text{V} = 0$$
$$V_A = 2.4\,\text{V}$$

Then we could use a voltage divider on the left branch of the circuit:

$$V_x = V_A \cdot \frac{2\,\text{k}\Omega}{2\,\text{k}\Omega + 2\,\text{k}\Omega} = \frac{V_A}{2} = 1.2\,\text{V}$$

However, since we have two independent sources, we can also use superposition to solve this circuit. Let's start by finding the contribution of V_x that comes from the voltage source. First, we'll zero out all other independent sources besides the voltage source of interest, which in this case just means zeroing out the 2 mA current source to create the circuit shown in Figure 3.21.

We'll ignore the resistor on the right side since it's floating and no current can flow through it. We can find Vx by using a voltage divider:

$$V_{x,v} = 5\,\text{V} \cdot \frac{2\,\text{k}\Omega}{2\,\text{k}\Omega + 2\,\text{k}\Omega + 1\,\text{k}\Omega} = 2\,\text{V}$$

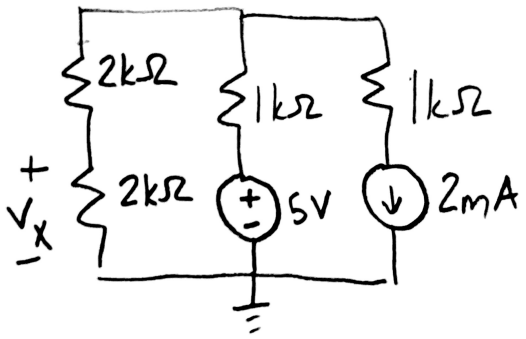


Figure 3.20: A circuit only a textbook would love. Before we solve this circuit, there is something strange about it. Can you spot it? Since devices in series have to have the same current, adding any device in series with a current source doesn't really do anything, so you never see this configuration in a real circuit.

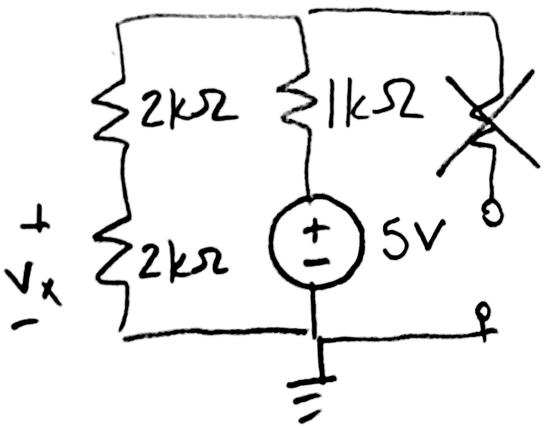


Figure 3.21: Circuit of Figure 3.20 with the current source set to zero. v_x can be found from a voltage divider, and is 2/5 of 5V, or 2V.

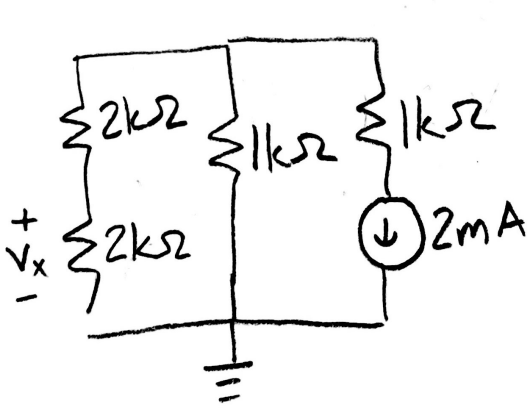


Figure 3.22: Circuit of Figure 3.20 with the voltage source set to zero. The current flowing through v_x can be found from a current divider, and is 1/5 of -2mA, or -0.4ma. Using Ohm's law, this means that v_x is -0.8V.

Now we can solve for the contribution from the current source. First, we'll zero out the voltage source by replacing it with a wire, yielding the circuit shown in Figure 3.20. Since the right 1K resistor is in series with the current source, it does nothing, and 2mA must flow through this resistor. Since the current in the current source is flowing down, the current flow in the resistors on the right part of the schematic must be up, which is opposite to the reference direction. Thus the current through these resistors will be negative. There are multiple ways to solve this, but we'll use a current divider to find the current through the $2\,\text{k}\Omega$ resistor, then multiply it by the resistance to find V_x. One branch in our current divider has a resistance of $1\,\text{k}\Omega$. The second branch has a resistance of $4\,\text{k}\Omega$, since we have two $2\,\text{k}\Omega$ resistors in series.

$$I_r = -2\,\text{mA} \cdot \frac{1\,\text{k}\Omega}{1\,\text{k}\Omega + 4\,\text{k}\Omega} = \frac{-2\,\text{mA}}{5} = -0.4\,\text{mA}$$

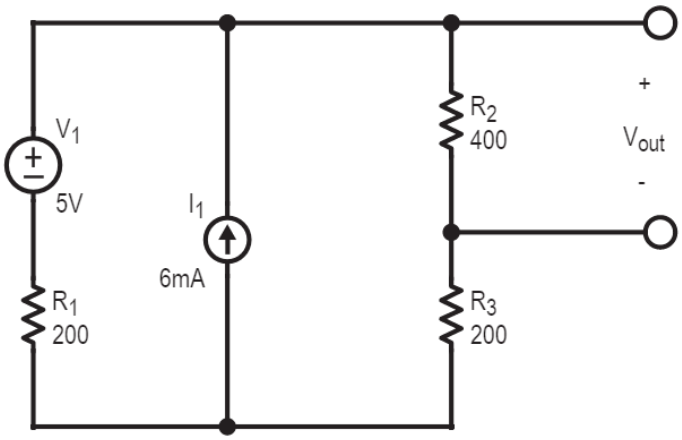
$$V_{x,i} = -0.4\,\text{mA} \cdot 2\,\text{k}\Omega = -0.8\,\text{V}$$

Now that we have the contributions to V_x from both independent sources, we can simply add them together to find the total V_x:

$$V_x = V_{x,v} + V_{x,i} = 2\,\text{V} + (-0.8\,\text{V}) = 1.2\,\text{V}$$

Note that this is the same answer that we would get from nodal analysis.

Problem 3.5
Find the output voltage of the following circuit using superposition.



3.6 Equivalent Circuits

(This section is optional reading for Engr 40M)

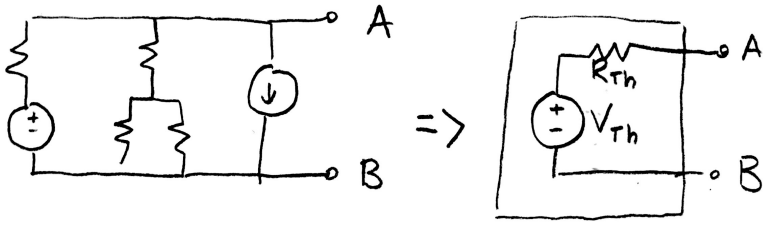
Honestly, most of the circuits that you will need to deal with when you are making stuff (and not doing homework problems) will be pretty simple, and won't be hard to analyze. Nodal analysis will always work, and you can make this even easier by reducing the complexity by using series parallel reductions and using current and voltage dividers. But occasionally you will find a circuit that is hard to analyze, or have a design problem where you will need to find a component that makes a voltage or current somewhere else in the circuit to the right value. Here it would be great if you could simplify the rest of the circuit to yield something simple like a voltage divider, when the relationship between the device you can adjust and the value you are interested in is easy to understand. This is especially true when there are some non-linear elements in the circuit, where understanding the region of operation of the non-linear device is important. Fortunately it is almost always possible to create these type of simplifications.

The basic idea behind these simplification is a concept we have used before: creating the iV curve for an electric device. In the previous chapter we created these curves for primitive devices like voltage sources, resistors and diodes, but we also showed how we could model more complicated device like a battery as a combination of a resistor and a voltage source, or a solar cell as a diode in parallel with a current source dependent on the light intensity. What we show in this section is that any combination of linear resistors, voltage sources and current sources can be modelled by a single resistor in series with a single voltage source. This single resistor, single voltage source model is called the *Thevenin Equivalent circuit*. The iV of any complex linear circuit can also be modeled by a current source in parallel with a resistor, and this model is called the *Norton Equivalent* circuit. The reason this is possible is simple. If all our devices are have a linear relationship between current and voltage (true for resistors, voltage and current sources) then the resulting relationship between current and voltage of any combinations of these devices will also be linear. This is the principle that allows us to use superposition. But it also means that if we take any two nodes in a linear circuit and pretend that all the components form a new device (we are just going to model the relationship between current through and the voltage across the two nodes that we selected), the relationship between current and voltage of this new "device" will be a straight line in the i-V plane.

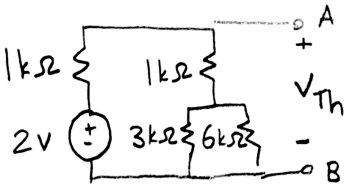
That line can always be fit by the equation $V = i \cdot R + V_{th}$, for the right choice of R (which is a resistance) and V_{th} which is a voltage. This line also be models the current voltage relationship of a circuit consisting of a voltage source with a value V_{th} in series with a resistor with value R.

3.6.1 Thevenin Equivalent

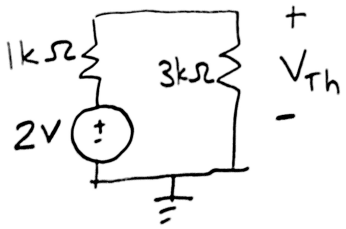
Thevenin's theorem states that a linear two-terminal circuit can be replaced by an equivalent circuit composed of a voltage source and series resistor, as shown below:



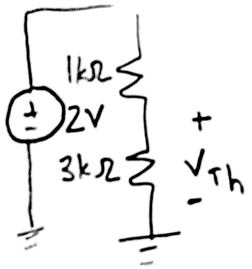
Further the theorem states that you can't make any measurement using just the nodes A and B that can tell the difference between the two circuits. this give us a hint of how to find the value of the voltage source and resistance for the equivalent circuit. One simple measurement we can make is to measure the voltage between A and B with nothing connected between the two terminals. In the Thevenin circuit, this voltage is obvious. Since there is no current running between A and B, the voltage drop across the resistor must be zero ($V = i \cdot R$), so the measured voltage is just V_{th}. Remember that since no current is flowing between the terminals, this is the "open-circuit" voltage. Since the open circuit voltage is equal to the value of the voltage source in the equivalent circuit, it is often called the Thevenin voltage. We can find this voltage by using nodal analysis on the actual circuit with no connection between A and B.



Let find the Thevenin equivalent voltage of the circuit shown above. We could use nodal analysis, and have two voltages to solve for, but in this case we can also solve the problem using series parallel reductions, so let's do it that way. We can immediately see that the 3kΩ and 6kΩ resistors are in parallel, so we can combine them into one resistance: $(\frac{1}{6k\Omega} + \frac{1}{3k\Omega})^{-1} = 2k\Omega$. This 2kΩ resistor is in series with the 1kΩ resistor above it, so we can add those together to form a single 3kΩ resistor. We end up with the circuit below:

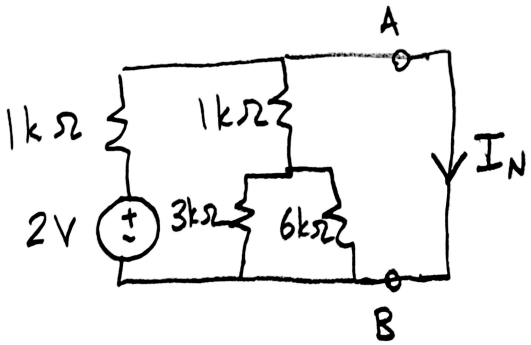


Here, we want to find the voltage across one of two resistors. This looks like a voltage divider. We can re-draw it to make that relationship clearer:

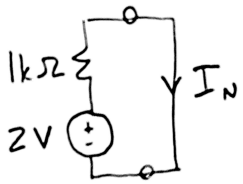


Now we can use our voltage divider equation, $V_{Th} = V_{in}\frac{3k\Omega}{3k\Omega+1k\Omega}$, to solve for V_{Th}. We end up with $V_{Th} = 1.5V$.

Now we need to make another measurement so we can figure out what the equivalent series resistor should be. Another easy measurement to make is the current that flows when we short together nodes A and B. This is the short circuit current.

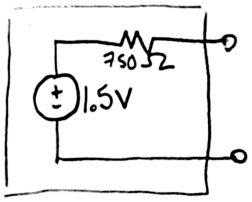


Remember the current through any resistor where both terminals of the resistor connect to the same node is zero (there is no voltage across the resistor so the current must be zero). This means that the 1 kΩ, 3 kΩ and 6 kΩ resistors can be removed, since there is no voltage across them.

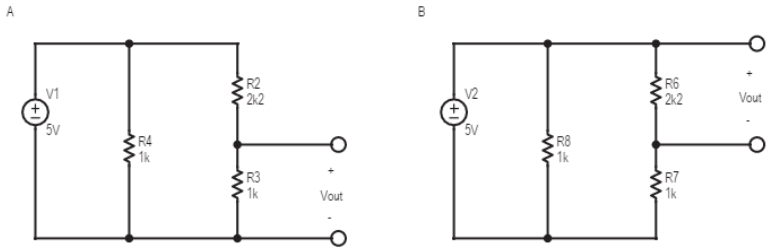


Finding the current in this case is easy, since we can just use Ohm's Law: $I_N = \frac{V}{R} = \frac{2V}{1k\Omega} = 2mA$.

To find the equivalent resistor we know that if we short the output pins of our equivalent circuit, we must also get 2 mA. So 1.5 V/R must 2 mA, so R = 750 Ω. Now that we have the Thevenin voltage and resistance, we can plug them back in to our black box model to come up with the following Thevenin equivalent:



Problem 3.6 : Thevenin equivalent
Find the Thevenin equivalents of circuit A and circuit B below, at the ports indicated by V_{out}. Try to apply the rules you've learnt in this chapter.



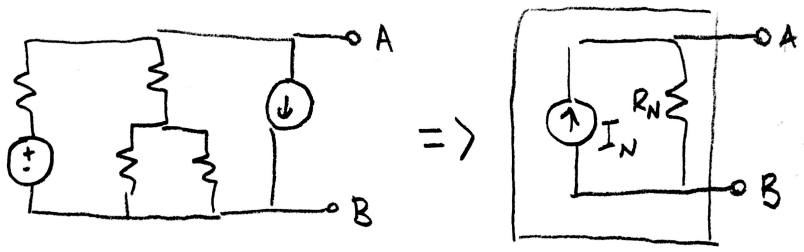
3.6.2 Norton Equivalent

It turns out that any linear i-V curve can also be represented by a current source in parallel with resistor, in addition to the Thevenin equivalent that we just described. Previously we said that any

linear circuit could be represented by $V = i \cdot R_{Th} + V_{Th}$, where R_{Th} is the Thevenin equivalent resistance and V_{Th} is the Thevenin equivalent voltage. But dividing both sides of this equation by R, and rearranging terms gives:

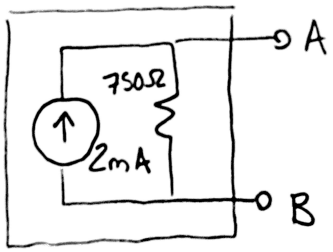
$$i = \frac{V}{R} + \frac{V_{Th}}{R_{Th}}$$

Since the current is the sum of two terms, it could be modeled by two elements in parallel. One is a resistor, since the current is proportional to a voltage, and the other term is a constant current, which can be modeled as a current source. Thus we get **Norton's theorem** which states that a linear two-terminal circuit can be replaced by an equivalent circuit composed of a current source and parallel resistor, as shown below:

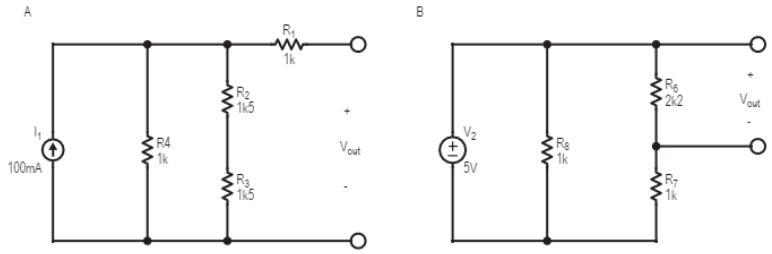


We solve this in a very similar manner to solving for Thevenin circuits. In this case the value of the current source is just the value of the short-circuit current, since with no voltage across the terminals, the current through the resistance must be zero. We found this before to be 2 mA.

Since the Norton resistance is equal to the Thevenin resistance, we use the same value we found before, 750 Ω.

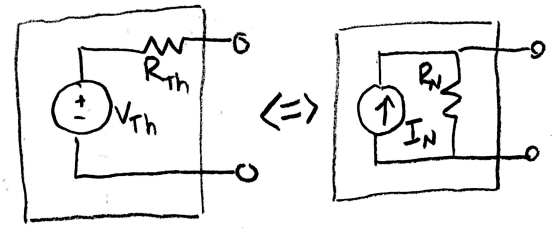


Problem 3.7 : Norton equivalents
Find the Norton equivalent of the following circuits.



3.6.3 Converting from Thevenin to Norton

As we saw above, the Thevenin and Norton resistances are equivalent. Thevenin voltage and Norton current can also be related using Ohm's Law:

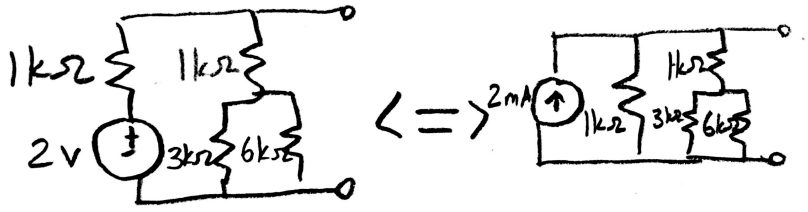


$$R_{Th} = R_N$$

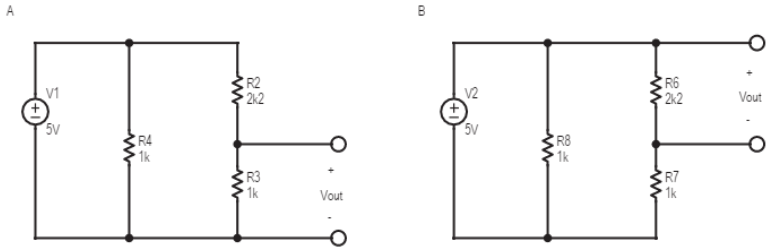
$$V_{Th} = I_N R_N$$

$$I_N = \frac{V_{Th}}{R_{Th}}$$

This result can be applied more generally to convert a voltage source in series with a resistor into a current source in parallel with a resistor, and vice versa. We can use this to simplify our nodal analysis. For example, if we wanted to work with a current source rather than a voltage source in the circuit we solved earlier, we could solve for $I_N = \frac{2V}{1k\Omega} = 2mA$ and update our circuit as shown below:

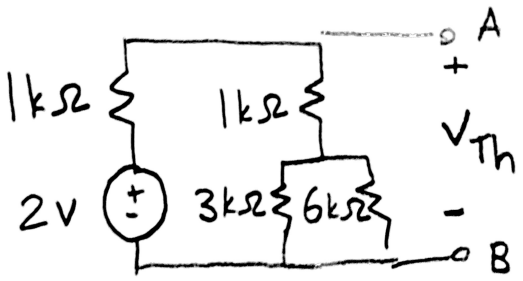


Problem 3.8
Converting from Thevenin to Norton:
Find the Norton equivalent of the following circuits - you derived their Thevenin equivalents earlier.
Hint: The Thevenin equivalent resistance of a circuit is equal to the Norton equivalent resistance of the circuit.

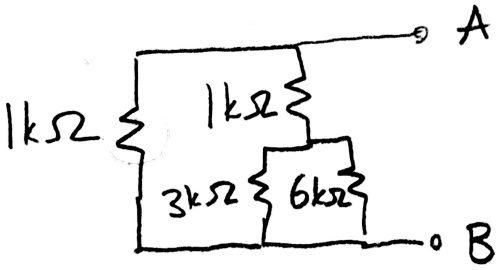


3.6.4 Using Superposition to Find R

Superposition also provides another method to solve for the Thevenin/Norton equivalent resistance. Because of superposition the actual circuit and the equivalent should still be the same if we set all sources to zero. The Thevenin equivalent circuit becomes just a single resistor, and the real circuit becomes a collection of resistors. Going back to the circuit we worked on before:



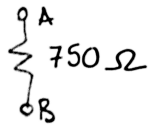
To solve for the equivalent resistance between terminals A and B, we first need to set all independent sources to zero. We want zero voltage across the voltage sources, which means we need to short them (replace them with a wire). We want no current across the current source, which means we need to open them (take them out and leave the two nodes unconnected). In our circuit, we only have one voltage source, which we'll replace with a straight wire. Once that's done, our circuit looks like this:



If we'd had a current source in place of the voltage source, we would cut it out, leaving the $1\,\text{k}\Omega$ resistor unconnected to anything. A dangling resistor makes no difference in the circuit since current can only flow through a loop, so we would just ignore the resistor, as shown below:

Let's return to our original circuit. We can simplify the right branch just as we did when finding the Thevenin voltage, giving us this circuit:

This is just two resistors in parallel. We can easily combine these resistors to form a single resistor, as shown below:



3.7 Congratulations!

You have now completed the first module in learning how to make stuff—understanding how electrical circuits work. For those of you who have never thought about voltage or current before, we have covered a large amount of material. We first introduced the notion of charged particles, called moving charged particles current, and the potential energy different of these charged particles voltage. We learned that charge is always conserved, and all nodes are charge neutral, so that means that the charge flowing into a device or node must always equal to the charge flowing out. We also learned about energy conservation which means that the sum of the voltages across devices in any loop is zero. We then used these rules to allow us to define node voltages, and a set of rules (nodal analysis) which allows us to find the node voltages and device currents for any circuit that we come across.

We have also learned about special constraints for circuit types of circuits. When two devices share exclusively one node, they must have the same current, and we can use this common current to help simplify our circuit analysis. We say that these devices are in series. We also know that when two devices share the same two nodes they have the same voltage across them, and again provides an opportunity for circuit simplification. We say these two devices are in parallel. Finally we know if the circuit only consists of resistors and voltage and current sources, the circuit is linear and superposition holds. This means we can evaluate the effect of each source independently.

Through the rest of this book we will continually rely on this basic knowledge to help you *make* even better stuff. The next chapter will look about how a small number of electronic devices have changed the world you live in, and help you make a silly toy.

3.8 Solutions to Practice Problems

Solution 3.1:
We can always choose one of the nodes in a circuit to be the "reference", or "ground" to simplify the problem. In this case we choose the bottom node for simplicity. Then, there is only one voltage node of interest we need to solve for - the node connecting R_1, R_2 and R_3 together. Let's call this node V_x.

By KCL, we know that $I_1 + I_2 + I_3 = 0$. Then we can write the following questions and solve for V_x:

$$\begin{aligned} I_1 + I_2 + I_3 &= 0 \\ \frac{V_x - 1_2}{1\text{k}} + \frac{V_x}{2.2\text{k}} + \frac{V_x - 24}{2.2\text{k}} &= 0 \\ 2.2 \cdot V_x - 2.2(12) + V_x + V_x - 24 &= 0 \end{aligned}$$

$$4.2V_x = 50.4$$

$$V_x = 12\,\text{V}$$

$$I_1 = \frac{12 - 12}{1\text{k}} = 0\,\text{A}$$

$$I_2 = \frac{12}{2.2\text{k}} = 5.45\,\text{mA}$$

$$I_3 = \frac{12 - 24}{2.2\text{k}} = -5.45\,\text{mA}$$

Solution 3.2:

$$R_{eq} = R_7 + R_6 \parallel (R_5 + R_4 \parallel (R_3 + R_1 \parallel (R_2 + R_8)))$$

Make sure you understand how this equation came about. It simply summarises the series and parallel relationships in the circuit.

$$R_{eq} = 1\text{k} + 1\text{k} \parallel (1\text{k} + 1\text{k} \parallel (1\text{k} + 1\text{k} \parallel (1\text{k} + 1\text{k})))$$

$$R_{eq} = 1\text{k} + 1\text{k} \parallel (1\text{k} + 1\text{k} \parallel (1\text{k} + 1\text{k} \parallel 2\text{k}))$$

$$R_{eq} = 1\text{k} + 1\text{k} \parallel \left(1\text{k} + 1\text{k} \parallel \left(1\text{k} + \frac{2}{3}\text{k}\right)\right)$$

$$R_{eq} = 1\text{k} + 1\text{k} \parallel (1\text{k} + 1\text{k} \parallel (1.667\text{k}))$$

$$R_{eq} = 1\text{k} + 1\text{k} \parallel 1.625\text{k}$$

$$R_{eq} = 1.619\,\text{k}\Omega$$

Solution 3.3:
Circuit A:

$$V_{out} = 12 \cdot \frac{R_2}{R_1 + R_2}$$

$$V_{out} = 12 \cdot \frac{2.2\text{k}}{1\text{k} + 2.2\text{k}} = 8.25\,\text{V}$$

Circuit B:

$$V_{out} = 12 \cdot \frac{R_3}{R_3 + R_4}$$

$$V_{out} = 12 \cdot \frac{1\text{k}}{1\text{k} + 2.2\text{k}} = 3.75\,\text{V}$$

Solution 3.4:
Circuit A:

$$I_1 = 100\,\text{mA} \cdot \frac{2.2\text{k}}{1\text{k} + 2.2\text{k}} = 68.75\,\text{mA}$$

$$I_2 = 100\,\text{mA} \cdot \frac{1\text{k}}{1\text{k} + 2.2\text{k}} = 31.25\,\text{mA}$$

Circuit B: The current divider equation doesn't scale nicely to more branches. So, when you see a circuit like this, it is easier to just use nodal analysis. First, find the voltage across all the resistors:

$$V = 100\,\text{mA} \cdot (1\text{k} \parallel 2.2\text{k} \parallel 4.7\text{k}) = 100\,\text{mA} \cdot 600 = 60V$$

$$I_3 = \frac{60}{1\text{k}} = 60\,\text{mA}$$

$$I_4 = \frac{60}{2.2\text{k}} = 27.27\,\text{mA}$$

$$I_5 = \frac{60}{4.7\text{k}} = 12.76\,\text{mA}$$

Solution 3.5:
The general procedure is as follows:
First calculate the contribution of V1 to the output voltage (call this V_{out1}), by removing the current source (replacing it with an open circuit).
Then, calculate the contribution of I1 to the output voltage (call this V_{out2}) by removing the voltage source (replacing it with a short circuit).
Finally, add the two voltages together to get V_{out}.

$$V_{out1} = 5\,\text{V} \cdot \frac{R2}{R1 + R3} = 2.5\,\text{V}$$

$$V_{out2} = V_{isource} \cdot \frac{R2}{R2 + R3} = 6\,\text{mA} \cdot R1 \parallel (R2 + R3) \cdot \frac{R2}{R2 + R3}$$

$$= 0.9\,\text{V} \cdot \frac{400}{600} = 0.6\,\text{V}$$

$$V_{out} = V_{out1} + V_{out2} = 3.1\,\text{V}$$

Solution 3.6:
Circuit A:

$$R_{TH} = (R_2) \parallel R_3 = 690\,\text{k}\Omega$$

$$V_{TH} = 5 \cdot \frac{1\text{k}}{2\text{k}2 + 1\text{k}} = 1.56\,\text{V}$$

Circuit B:

$$R_{TH} = (R_7) \parallel R_6 = 690\,\Omega$$

$$V_{TH} = 5 \cdot \frac{2.2\text{k}}{2.2\text{k} + 1\text{k}} = 3.44\,\text{V}$$

Solution 3.7:
Circuit A:

$$R_N = R_1 + R_4 \parallel (R_2 + R_3) = 1.75\,\text{k}\Omega$$

To find I_N, we short the output ports and find the current flowing through them.
The voltage across the current source under those conditions are:

$$V_{current} = 100\,\text{mA} \cdot (R_4 \parallel (R_2 + R_3) \parallel R_1) = 100\,\text{mA} \cdot 429\Omega = 42.9\,\text{V}$$

$$I_N = \frac{V_{current}}{R_1} = 42.8\,\text{mA}$$

Circuit B:

$$R_N = (R_7) \parallel R_6 = 1.05\,\text{k}\Omega$$

$$I_N = \frac{5\,\text{V}}{1\text{k}} = 5\,\text{mA}$$

Solution 3.8:
Circuit A:

$$R_N = R_{TH} = 690\,\Omega$$
$$I_N = \frac{5}{2.2\text{k}} = 2.27\,\text{mA}$$

Circuit B:

$$R_N = R_{TH} = 690\,\Omega$$
$$I_N = \frac{5}{1\text{k}} = 5\,\text{mA}$$

Chapter 4

Introduction to Digital Logic

You are surrounded by the fruits of the information revolution, from the smart phones that you carry around to the web services that you depend upon. While this equipment seems very different from the material we studied in the first three chapters, it turns out that all these devices are built from electrical circuits, which can be analyzed by the procedures you have already learned. But this analysis often can be simplified, since these circuits usually work on signals which only have two legal values. These values are sometimes called true and false, and other times called 0 and 1. It is amazing to think that our complex phones/computers/web are built from large numbers of simple elements operating on these *binary* signals.

As an introduction to this world of 0s and 1s we will start with a very silly machine, called a useless box. This machine will show why some signals only have two useful values, and how we can compute with these values to do interesting operations.

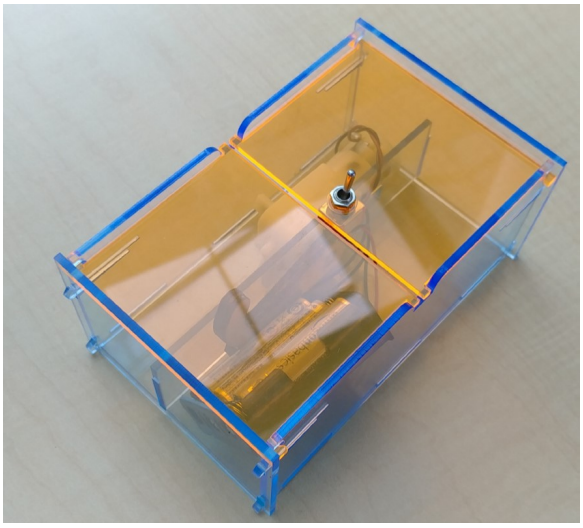


Figure 4.1: A useless box

4.1 Useless Box

The origins of this machine date back to the 1930s. Once turned on, the box's sole purpose is to perform the mostly useless task of turning itself back off. In a little more detailed description, the useless box is a machine with a simple two position switch which, when flipped to the on position, causes an internal lever to peek out from inside the box and flip the switch back to the off position, after which the lever retreats back inside the box. In its most simple implementation the box contains two switches (the flip switch that serves as the user interface, and a limit switch), a motor, and a battery pack. Smarter versions of the boxes, which can be programmed to execute their useless tasks in more creative ways, will also include a microcontroller of some sort, in our case an Arduino. You will build both versions of this box in lab, but this section will focus on the original, simple box. Before we can explain how the box works, we need to introduce two new electro-mechanical devices, a switch and a motor.

4.1.1 Switches

I am sure that you have used electrical switches many times, and the first chapter already mentioned them when it talked about a flashlight. But have you ever thought about how they work and why we need them? Switches are a way for people to use mechanical motion, which they are good at, to change the current flow in a circuit (which they are not that good at). Figure 4.2 shows a number of different shaped switches, from the "wall" switches you are all familiar with to other switch forms found in electrical systems. Some (the left two in the picture) have a couple stable positions, and will remain in their current state until someone "switches" them. Others (the right two in the figure) only make a connection when the button or lever is depressed, when the force is removed, the switch returns to its default state.

Not only do switches have different physical properties, their electrical connections can change as well. Switches are electrically characterized by the number of poles and throws they support. Poles are a measure of how many parallel circuits are switched by the switch and throws are a measure of how many different terminals the switch can possibly connect to. Figure 4.3 shows four example switches, each with a different electrical configuration. The single pole, single throw (SPST) switch has only two terminals, and either connects the two terminals, or leaves them not connected (open circuit). A double throw version of this switch has three terminals, and either connects the left terminal to the top right or the bottom right terminal. When the top right terminal is connected, the bottom right terminal is left unconnected—it is an open circuit. A number of these single pole switches can be mechanically linked together to form a multiple pole switch, as shown on the bottom row of the figure. In switches with multiple poles, there is no electrical connections between the different switches ganged together. The linkage is only mechanical, forcing the two switches to always have the same connection pattern as each other. So for a DPST switch, the two SPST switches that are linked together would either both be closed (connected) or both would be open circuits.

In the useless box we will use switches for two purposes. First we need a switch for the user to control. As we will see later, this switch needs to be a DPDT switch. But we will need an additional switch to make the box function, since we need to know when the "finger" is fully retracted into the box. We will use another switch to do that, and have the finger push on a momentary contact switch when it is fully retracted inside of the box.

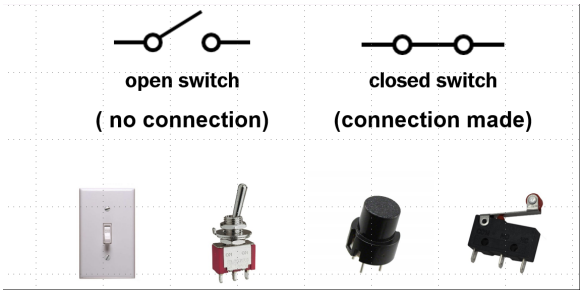


Figure 4.2: Switches come in many shapes and forms. Some switches have many stable positions, others only connect when you push them, and return to their default position. But the function of all switches is the same: to make and break electrical connections between terminals.

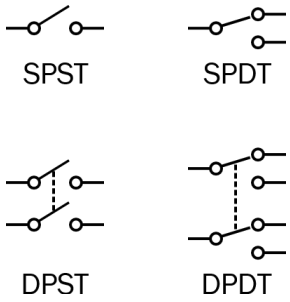
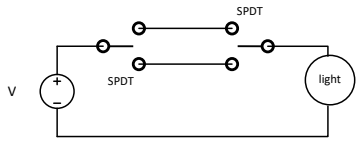


Figure 4.3: Switches are characterized by the number of throws (the number of terminals that a switch can possibly connect to) and poles (the number of parallel electrical switches that are mechanically linked together). The mechanical linkage is shown by the dotted line, and indicates that the two switches always have the same configuration.

Question: How do lights work which have two light switches?

We all have used lights where there are two switches that each control the light. With either switch you can turn the light on or off by changing the position of the switch. Have you ever thought about how that circuit works? With your new circuit model you can work it out. The circuit requires two SPDT switches. The common terminal of one switch goes to the power supply. The common terminal of the other switch goes to the light. The other end of the light goes to the other side of the power supply. This is shown below:



To get the behavior we want, now we only need to connect the top terminal of one switch to the top terminal of the other switch, and the bottom terminals together as well. If both switches are on, the light is on, and if both are off, the light is also on. If they don't match, the light is off. Usually rooms are wired so that the light is on if only one switch is up. How could you change the circuit so that the light is on when one switch is up and one is down? (Thus the "off" state is both switches down or both up.)

4.1.2 Motors

Not only does the box need a way to convert mechanical input into electrical signals, it also needs a way to convert electrical signals into mechanical action. This conversion is often done by a motor. Motors convert electrical energy into mechanical energy, and can be run backward to convert mechanical energy into electrical energy. The design and operation of motors is a fascinating area, but one that we don't have time to cover in this class. There are a number of good tutorials about motors on the web if you are interested in learning more about them. I would start off with https://learn.sparkfun.com/tutorials/motors-and-selecting-the-right-one if you are interested.

You will be using a DC brush motor in your project, as shown in Figure 4.4. The two wire terminals of the motor run through coils of wire on the armature (the moving part of the motor) that form electromagnets as shown in the middle picture in the figure. There is a lot of wire in the motor, so even though the wire resistance is small, it adds up. The total resistance of the wire depends on the motor. For the motors that we use in the class, it is around $10\,\Omega$. This means if the motor is not turning, the motor will look like a $10\,\Omega$ resistor. But if the motor starts to turn, its characteristics change dramatically. A voltage is created across the coiled wire, which is a physics phenomenon we will talk more about in Chapter 9. The motor now can be modeled by a voltage source in series with the resistor, where the value of the voltage source is proportional to the speed that the motor is turning. This motor model is shown in the right diagram in the figure.

A motor converts electrical energy to mechanical energy. The voltage across a motor sets how fast the motor turns. An ideal motor (with no friction, and driving no load) would not require any current once it was spinning at the right rate. Its internal voltage source would match the voltage supplied by the voltage source, and no current would flow. If the motor would need to do some

work (it needed to drive a load, or it had friction it needs to overcome), then current would be required. The amount of current depends on how hard the motor needs to work (push) to continue to turn. The power going into the motor (its current times the voltage across it) must be larger than the mechanical work that the motor does (conservation of energy). If the wire resistance was zero, then all the energy would be transferred. Since the resistance is not zero, some of the energy is lost in heating up the wire. This simple model explains why the speed of a motor depends on the drive voltage and the mechanical load. Under small load (low current) the voltage across the resistor should be small, and the motor voltage source would be similar to the voltage driving the motor. As the mechanical load increases, more current is required, and the voltage drop across the resistor increases, so the voltage source in the motor needs to be smaller, and the motor slows down.

This model of a motor also explains how a motor can be used to generate electrical power. If you manually spin a motor, the internal voltage source will increase in value, and you can measure this voltage at the pins of the motor. The more current you pull out of the motor (by having it drive a smaller resistor), the harder it will be to turn the motor. If you don't believe me, you should try it.

Take the finger for your useless box and put it on your motor. This should allow you to turn the motor by hand. Connect your DMM across your motor and measure the voltage the motor produces. This measurement takes little current from the motor, and you should find the motor easy to turn. You will also see that the faster you turn the motor, the higher the voltage you generate. Turning the motor in the opposite direction will generate negative voltage. If you now switch your meter from voltage mode, to measuring current, the meter now puts a small resistor across the terminals of the motor, requiring the motor to provide a large current to raise the voltage. You will find turning the motor now is much harder!

4.1.3 The "Brains" of a Useless Box

Now that we understand how our new electrical devices work, we can use them to create the "brains" of the useless box. To help understand what type of switch we need to use, and how we should connect them up, we can create an abstract model of the box called a *finite state machine, or FSM*. An FSM model works well when the behavior of the system can be described as transitions between a modest number of operating conditions (or states). The useless box is one of these types of machines.

Its starting condition is with the motor off, and the switch in the "off" position. Let's call this the "off" state. The box will stay in this state until someone flips the control switch into the "on" position. Now the motor starts, and moves the finger out of the box to turn the switch off. Let's call this operating state, "forward". The motor continues to turn pushing the finger out of the box until the finger hits the switch, putting it back in the off position. At this point we want the motor to change direction and pull the finger back into the box. In this operating state, the motor is on, but moving in the opposite direction it was moving before. Let's call this state "reverse." In this state if the user flips the switch to "on" when the finger is retracting, we want it to transition to the "forward" state, and move to flip the switch to "off" again.

Now we have a dilemma. How does the box know when to transition from the reverse state back to the off state? We would like this transition to happen (and for the motor to turn off) when the finger is fully retracted into the box. We can detect this situation by using an additional switch

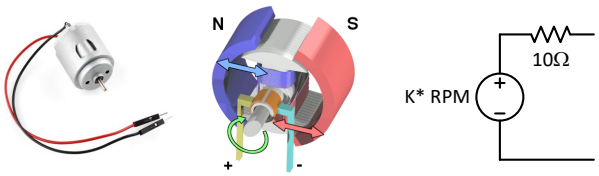


Figure 4.4: A picture of a DC brush motor, which is the most common type of motor, a schematic of how it works, and a circuit model of the motor. A brush motor traditionally consists of two parts, a fixed magnetic field and a changeable field from an electromagnet. The fixed magnetic field is often a permanent magnet outside of the rotating shaft. Current flowing through the motor energizes electromagnets attached to the inside rotating section, called the armature, causing them to also generate a magnetic field. The force of the opposing magnetic fields cause the armature to rotate and align the fields. As the motor turns, the connections to the coils generating the magnetic field change, causing a new electromagnetic to turn on, which again tries to align with the external field. When this coil is near alignment, the current switches again and the process repeats, causing the armature to continue to turn. A motor can be nicely modeled as a voltage source in series with a resistor. The value of the voltage source is proportional the how fast it is spinning, and the value of the resistance is the sum of all the resistance in the coils of the motor.

placed under the finger in the bottom of the box. When the finger retracts far enough to hit this switch, we will transition from the reverse state to the off state. So the transition from the reverse state to the off state happens when the "limit" switch is hit. For this switch to work as desired, we need this limit switch, to be a momentary contact type of switch. That is we want the switch to change state as soon as the finger is no longer depressing it.

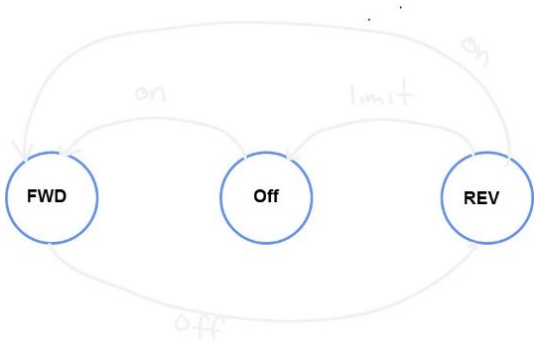


Figure 4.5: The FSM diagram of a useless box. The states are represented by circles and the transitions are by faint arcs.

This entire control scheme can be represented by a FSM diagram, which displays the states as circuits, and displays the transitions between states as arcs. This FSM diagram is shown in Figure 4.5. Although fairly simple, the FSM diagram provides a powerful framework for the design of a digital system or a program that needs to keep track of where it is in a complex sequence of tasks. Its importance is why it has a fancy name: finite state machine or FSM. If you take more

hardware design classes, like EE108, you will learn much more about FSM. But we will end this short discussion of FSM to look at little more closely at how to connect your switches to implement this FSM.

One of the challenges that we first face in trying to wire up the two switches, battery pack and motor is that we need to have the motor run in two different directions. Let's label the terminals of the motor **M+** and **M-**. If the positive end of the battery is connected to **M+**, and the negative end of the battery is connected to **M-**, the motor is in the "forward" state, and pushes the finger out of the box. To retract the finger, we need to reverse the battery connections, with the negative end of the battery connecting to **M+**. To reverse the motor, we need to switch the battery connection on both wires, and so the "on-off" switch needs to be a DPDT switch. Here we can connect **M+** to the common terminal of one pole (one part of the switch) and connect **M-** to the common terminal of the other pole.[1] In the "on" position, we connect the positive end of the battery to the **M+** pole, and connect the negative end of the battery to the **M-** pole of the switch.

The connections for the "off" position are a little more complex, since we need to turn the motor on only if the limit switch is **not** depressed. We can accomplish this by putting the limit switch in series with the battery connection (either the positive or negative since current must flow in a loop) before connecting it to the terminals of the "on-off" switch.

Once we have wired up the switches and battery, we can measure the voltages on the motor's terminals in each of its states. The results are shown in the table below. Depending on which battery lead is interrupted with the limit switch, the off state can be either with with 0V or 4.5V on both leads. Notice that the voltage on the motor terminals has only two stable values. It is either at the voltage of the reference value, which we call Gnd, or it is at the voltage of the battery pack. For historical reasons, this voltage is often called Vdd, and we will use this name as well. Since the voltages only have two values, they can be thought of as a Boolean signal, which is described in the next section.

State	M+	M-
Forward	4.5V	0V
Reverse	0V	4.5V
Off	0V	0V
or	4.5V	4.5V

4.2 Boolean Signals

While the useless box is kind of a silly toy, it serves as an introduction to a very powerful idea that has shaped the world that we live in today: digital logic. It is digital logic in small chips in your computers, phones, TVs, networking devices, and cloud servers that provide the services that we all have come to expect. To introduce you to this world, we will continue with the example of the useless box. In the world Boolean signals and digital logic, the signals to the motor are two Boolean signals, and the control of the box can be expressed by two simple equations.

[1] You can also have the battery connected to the common switch terminals and generate a different, but correct circuit.

A Boolean signal is one that has only two values, true or false. If we are going to represent a Boolean signal in an electrical circuit, we can use voltage to indicate the value of the signal. Generally we use the following mapping:

TRUE	1	HIGH	4.5V (Vdd)
FALSE	0	LOW	0V (Gnd)

What we have just created is a way to map an electrical voltage to a logical value. As in all electrical circuits, the way you find the voltage is through nodal analysis, or some analysis short-cut, but once you do this analysis you will find that the resulting voltage will either be very close to the power supply (Vdd) or have a value that it very close the the reference, Gnd. Since node voltage generally will only have one of two values, we say it is a Boolean signal, and contains one bit of information.

Let's look again at the voltages found on the motor's terminals. Using this mapping of voltages to logical values, we get the following table to the right. Notice that with this electrical mapping, we can say that the motor moves forward when **M+** is 1 (true) , and **M-** is 0 (false). Similarly if **M-** is 1 and **M+** is 0, the motor runs in reverse. In this binary view, it makes sense to call the **M+** wire Forward, and the **M-** wire Reverse, since when one is true and the other is false, the motor will move in the direction of the signal that is 1.

State	M+	M-
Forward	True	False
Reverse	False	True
Off	False	False

When we look at the useless box circuit a little further, we see that the voltage on the switches also form electrical binary signals. The voltages at the output of these switches will also only be Gnd or Vdd, so they are binary. This means that we can represent them in our circuit as a Boolean signal too. Let's represent the state of the two switches in the box by two Boolean signals, onSwitch, which represents the state of the on/off switch, and limitSwitch, which is the state of the limit switch. Now we can represent the operation of the useless box using simple Boolean expressions.

4.3 Boolean Operations

Boolean operations are operations involving only the values 1 (also known as `true`) and 0 (also known as `false`). The three basic operations are `NOT` (`!`), `AND` (`&&`) and `OR` (`||`).

These are most easily described with tables exhaustively listing all combinations, known as *truth tables*; these are listed in Table 4.1.

The names of the operations are clearer if we observe the following:

- `NOT(X)` or `!X` is the opposite of whatever `X` is.

Table 4.1: Truth tables for `NOT` (`!`), `AND` (`&&`) and `OR` (`||`) operators

`NOT` (negation)

`A`	`!A`
0	1
1	0

`AND` (conjunction)

`A`	`B`	`A && B`
0	0	0
0	1	0
1	0	0
1	1	1

`OR` (disjunction)

`A`	`B`	`A || B`
0	0	0
0	1	1
1	0	1
1	1	1

Table 4.2: Truth tables for `NAND`, `NOR` and `XOR` operators

`NAND`

`A`	`B`	`NAND(A,B)`
0	0	1
0	1	1
1	0	1
1	1	0

`NOR`

`A`	`B`	`NOR(A,B)`
0	0	1
0	1	0
1	0	0
1	1	0

`XOR`

`A`	`B`	`XOR(A,B)`
0	0	0
0	1	1
1	0	1
1	1	0

- `AND(X,Y)` or `X && Y` is 1 if, and only if, *both* `X` *and* `Y` are 1.
- `OR(X,Y)` or `X || Y` is 1 if, and only if, *either* `X` *or* `Y` is 1.

It is also common to define `NAND`, `NOR` and `XOR`, as follows:

- `NAND(A,B)` (*not and*) is `NOT(AND(A,B))`; that is, first take `AND`, then invert the result.
- `NOR(A,B)` (*not or*) is `NOT(OR(A,B))`; that is, first take `OR`, then invert the result.
- `XOR(A,B)` (*exclusive or*) is 1 if *either* `A` *or* `B` are 1, *but not both*.

The truth tables of these three operators are given in Table 4.2. Note that `XOR(A,B)` can be expressed as follows: `XOR(A,B) = (A && !B) || (!A && B)`.

There is a branch of algebra, known as *Boolean algebra*, that studies how these operators combine and interact with each other—much like the study of how addition and multiplication interact with each other in normal algebra. Boolean algebra is intimately connected with propositional logic. We don't study Boolean algebra deeply in ENGR 40M; we'll mainly be concerned with how to *implement* these functions using electronic circuits.

We will now use these operators to define the behavior of our useless box. What combination of the switch states must occur in order for the motor to be going forward? For the box to start moving in the forward direction, the user has to have flipped the user switch, so we know that SwitchOn must be true. We dont want to look at the limit switch in this case, since we want to finger to move forward whether the finger is fully retracted (limitSwitch is true), or if it is still moving back into the box. What makes the box fun is that the finger reverses as soon as you flip the switch. Because of this, the only condition for Forward to be true is SwitchOn to also be true. The Boolean equation is therefore: **Forward = SwitchOn**.

Now, we ask the same question for the reverse direction. For the box to start moving in reverse, the user switch must have been flipped back again (either by the box's lever or by the user). The box moves in reverse until it has hit the limit switch, which in this case should shut off the box,

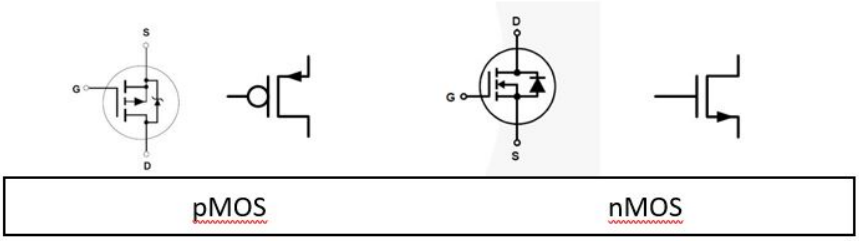


Figure 4.6: PMOS and NMOS symbols. In all the symbols, the gate terminal comes off to the left in the middle of the device. For the pMOS devices (on the left), the source is the terminal at the top of the device, and the drain is the terminal at the bottom of the device. For the nMOS devices (on the right) the source is the terminal on the bottom of the device, and the drain is the terminal on the top of the device.

otherwise the lever would forever be in reverse. This means that the box is only in reverse while both the user switch and the limit switch are off. The Boolean equation for the reverse motion is therefore: **Reverse = !SwitchOn && !Limit**. There is a curious and important characteristic of logic functions which makes it possible to express the same logical equation in a number of different ways. For example we can also write Reverse as being true anytime that neither the user switch or the limit switch are on: **Reverse = !(SwitchOn || Limit)**. If you are interested in learning more about how to find these equivalent forms, you should take a class in logic design like EE108.

In the case of the useless box, we have created this logic using mechanical switches and motors. While this approach does work, and it is fun to play with, each switch is pretty big, and they don't switch very rapidly. Wouldn't it be great if we had a new type of switch that we could control with another electrical signal? There is such a device, and it is called a MOS transistor.

4.4 MOS Transistors

A *MOS transistor* is a new type of electrical device which has three terminals.[2] It has many useful capabilities, and can be used to build amplifiers and Boolean logic gates. This section will focus on the operation of MOS transistors in building digital, Boolean circuits. If you are interested in other uses of MOS transistors, you should take a class in circuit design, like EE114. In many ways a MOS transistor acts like a single pole, single throw (SPST) mechanical switch, except that the control is provided by another electrical terminal, rather than a mechanical input. But like all things in life, there are some subtle issues one needs to understand in order to use a MOS transistor as a switch.

The first complexity is that there are two types of MOS transistors. One is called a nMOS transistor, and the other is called a pMOS transistor. Their names come from the fact that the charge carriers that flow in each transistor are different. Electrons flow in an nMOS. Since electrons have negative charge these are (n)MOS transistors. Holes, which are mobile positive charge carriers flow in pMOS devices, hence (p)MOS. Both type of transistors have three terminals. The terminal

[2]In some situations it is important to model a fourth MOS terminal, the substrate, but this terminal is not important for the circuits we will create in E40M

which controls the flow of current is called the gate (g). Current flows between the source (s) and the drain (d) terminals. So the gate terminal is like the mechanical lever that sets the switch's position, and the source and drain are like the two terminals of the SPST switch.

There are many symbols that are used to denote nMOS and pMOS transistors. Some use arrows to show direction of current flow, others use a bubble on the gate terminal to distinguish the pMOS device from the nMOS device. Some symbols for MOS transistors are shown in Figure 4.6. When an arrow is present on the lead, it always indicates the source, and the arrow points in the normal direction of current flow through the device. Since most circuit diagrams have Gnd at the bottom of their circuit, and Vdd (the power supply voltage) at the top of the circuit, current generally (but not always) flows from the top of the circuit to the bottom of the circuit. For this reason, the source of the pMOS devices are generally placed at the top of the transistor (the source of the positive charge), and the source of the nMOS devices are placed at the bottom of the transistor (the source of the negative charge). In both cases, current flows down through the transistor. Additional reasons for this choice will become clear soon.

4.4.1 Simple Switch Model

While the gate terminal controls the current through the transistor, there is no[3] current flow from the gate to either the source or drain terminal. In other words there is an open circuit connection between the gate and the other two terminals of the device. Thus if we look at both the gate-to-source and gate-to-drain i-V curves, shown in Figure 4.7, we see that there is no current flow through the gate terminal. The iV curves show an open circuit. These curves shown V_{gs} vs I_{gs}. Always remember that **no** current flows into the gate terminal of a nMOS or pMOS transistor.

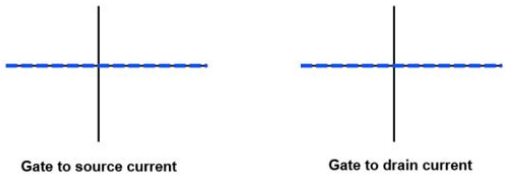


Figure 4.7: iV curve for Gate-Source terminals and Gate-Drain terminals

While no current every flows through the gate terminal, its potential is very important in determining the state of current flow between the other two terminals. We learned in Chapter 1 that electrical potential has to be measured between two nodes. So the potential of the gate can't control the switch (that is only one terminal). It is the voltage from the gate to the source terminal that determines the state of the MOS "switch." Remember the gate controls the conduction between the source and drain without having any current flowing into the gate terminal. How the gate to source voltage controls the conduction between the source and drain differs for nMOS and pMOS devices, so we will look at each in turn. We will first look at a simplified, idealized model, and then make it more realistic.

[3]Well modern transistors are so small that carriers can quantum mechanically tunnel through a material that doesn't conduct current and appear on the other side causing some current to flow, but we will ignore that issue in E40M. All of the transistors we will use won't have that problem.

Ideal switch model of nMOS

In an **nMOS** transistor, when the gate to source voltage, v_{GS}, is greater than the *threshold voltage* V_{th}, the transistor turns on and the "switch" connects the source and drain terminals. Otherwise, the transistor is off and the connection between the drain and source is an open circuit. Equivalent circuits in each of these cases are shown in Figure 4.8. The threshold voltage of an nMOS is typically between 0.4 V and 1 V.

nMOS off: when $v_{GS} < V_{th}$ nMOS on: when $v_{GS} > V_{th}$

Figure 4.8: Equivalent circuits for ideal switch model for nMOS. When the v_{GS} is small no current flows through the transistor. When v_{GS} is much larger than the threshold voltage, the source drain terminals are connected together.

We can combine these two cases into a single model by drawing a switch that closes when $v_{GS} > V_{th}$, as shown in Figure 4.9.

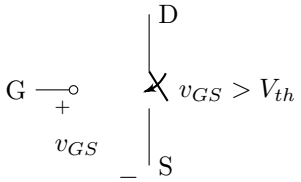


Figure 4.9: Ideal switch model for nMOS, showing how a nMOS transistor behaves like a SPST switch which is controlled by v_{GS}.

Ideal switch model of pMOS

The **pMOS** is similar, except that it's flipped: it turns on when $v_{GS} < V_{th}$, and V_{th} is negative, typically between -1 V and -0.4 V. Remember that V_{th} for a pMOS transistor is negative, so it turns on when the gate voltage is much *lower* than the source voltage. Equivalent circuits are shown in Figure 4.10.

pMOS off: when $v_{GS} > V_{th}$ pMOS on: when $v_{GS} < V_{th}$

Figure 4.10: Equivalent circuits for ideal switch model for pMOS.

Again, we can combine these cases into a single model, as in Figure 4.11.

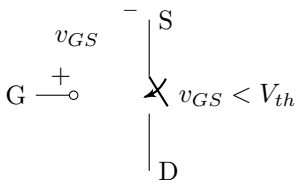


Figure 4.11: Ideal switch model for pMOS, showing how a pMOS transistor behaves like a SPST switch which is controlled by v_{GS}.

Resistance switch model of MOS

Real transistors have resistance when they are on. This resistance can be significant, so should be included in our model. This resistance can easily be modeled by adding a resistor in series with our switch. If we call the "on resistance" R_{on}, we get our final transistor models, shown in Figure 4.12 and 4.13. It's often easiest to think of a MOS transistor as a switch between the drain and the source, whose state is controlled by v_{GS}. If you see a transistor in a circuit, a good place to start is to ask yourself whether the switch is on or off, and then replace the transistor with the appropriate model shown in Figure 4.12 or 4.13.

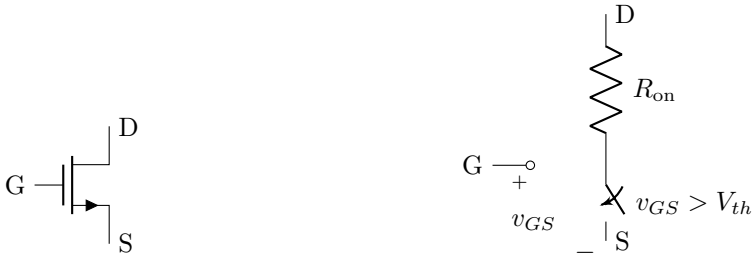


Figure 4.12: Internal resistance model for nMOS

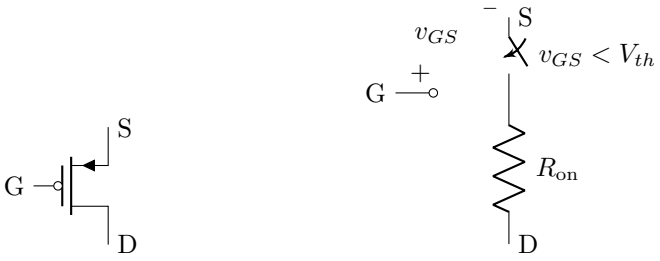
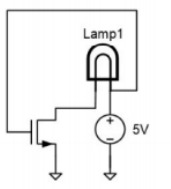


Figure 4.13: Internal resistance model for pMOS

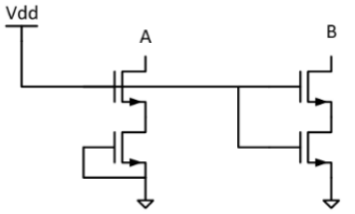
Problem 4.1 nMOS Transistor Behavior
Is the light in this circuit on or off?

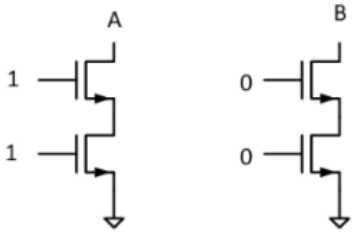


Problem 4.2 nMOS Transistor Characteristics
In the following circuits, what is:

- The resistance from A to Gnd?
- The resistance from B to Gnd?

Answer in terms of R_{on} of the transistors.

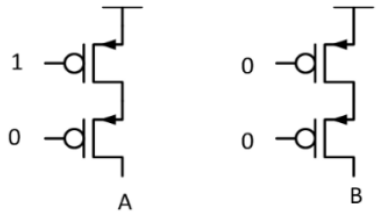




Problem 4.3 pMOS Transistor Characteristics
In the following circuits, what is:

- The resistance from A to Vdd?
- The resistance from B to Vdd?

Answer in terms of R_{on} of the transistors.



4.4.2 Using MOS transistors

While the previous section described how to create simple switched resistor circuit models for MOS transistors, it didn't really explain why we need both types. Since both pMOS and nMOS transistors act as switches, can't we do everything we need with just one type of transistor? The answer, not surprisingly is no, and this section describes why you need both types of transistors. It also will explain why the source of pMOS devices is at the top of the transistor, while the source of nMOS transistor is at the bottom, and why a pMOS transistor has a bubble (which usually means inversion) on its gate terminal.

While an nMOS transistor seems like it can do everything, it does have a serious limitation. It turns on when the gate to source voltage is much larger than a threshold voltage, and turns off when that voltage is small. This type of control is great if the source of the nMOS transistor is connected to a stable, unchanging voltage, like say Gnd. In fact, when the source of an nMOS transistor is connected to Gnd, and the signal on the gate is a Boolean signal, the binary value of the signal connected to the gate determines the state of the switch. If the signal feeding the gate terminal is a '1', then the voltage will be Vdd, and the gate to source voltage will also be Vdd, which is much greater than the threshold voltage. Thus the transistor will be on, and the MOS transistor will connect the drain to Gnd. When the input feeding the gate is a '0', the gate voltage will be Gnd, and the gate to source voltage will be 0V. The transistor will be off in this case, and the drain will be left floating (it is an open circuit). Thus when an nMOS's source is connected to Gnd, an nMOS transistor becomes a switch to Gnd where the state of the switch is set by the logical value on its gate terminal.

So an nMOS transistor is great for connecting a node to Gnd, or leaving it not connected. This is only half of what we need to build a logic gate. When the output line is driven to Gnd, we create a logical '0' value. To create a logical '1' value, we will also need to create a way to drive a signal to Vdd. At first it seems like we can simply use another nMOS transistor, and now connect its source to Vdd. But this doesn't work—to turn this transistor on we would need to create a gate voltage that is much larger than the source voltage. But the source voltage is already at the power supply voltage, which is usually the highest voltage in the circuit!

This is where pMOS transistors become useful. At first pMOS devices seems useless, since they turn on when their gate to source voltage is negative and in most logic circuits there aren't any negative voltages: their supply voltage is positive 1-5V above the reference. But the situation is not hopeless. Perhaps you have already figured out the trick. Remember that voltages are all relative. A voltage can be positive or negative depending on the reference directions that you use to measure it with. A pMOS device turns on when the gate to source voltage is much lower than its threshold voltage. If the source of the transistor is connected to Vdd, then any voltage between Vdd and Gnd will be a *negative* voltage! In fact, when the source of pMOS device is connected to Vdd, and its gate terminal is connected to a Boolean signal, it too acts like a switch where the state is

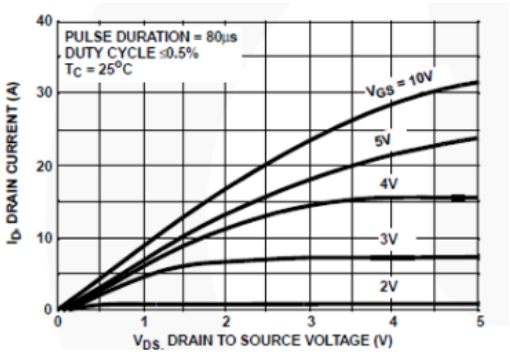


Figure 4.14: iV curve for Drain-Source terminals - V_{ds} vs I_{ds} for different applied V_{gs}

controlled by the value of the Boolean signal, with one twist. The switch connects the source and drain, connecting the drain to Vdd when the gate to source voltage is a large negative Voltage. This condition occurs when the signal on the gate terminal is at Gnd, or the Boolean signal is a '0'. When the gate terminal is at Vdd, or a logical '1', the gate to source voltage will be zero, and the transistor will be off. The pMOS transistor turns on when the input signal is *false*! That is the reason for the bubble on the gate input of a pMOS device. It is to remind you that the logical sense of the gate signal needs to be inverted to determine the state of the pMOS switch.

These properties of nMOS and pMOS transistor make it very easy to use them to create circuits that implement Boolean functions, which will be described in Section 4.5. Before we leave MOS transistors there is one more aspect about their behavior that you should know.

4.4.3 Real MOS Current Voltage Curves (deep background)

Our switched resistor model for a MOS transistor is a very useful model for the operation of a MOS transistor, and can be used to explain almost all of the issues that arise in creating digital logic gates, from predicting a gate's performance, to analyzing its power dissipation. But like all models of physical elements, it is an approximation of the actual behavior of the device. In particular this model works well when all the voltages are either near Vdd, or near Gnd, as they are in digital logic. This model is not accurate enough to use in circuits like an amplifier, where the voltages can be any potential between Vdd and Gnd.

Figure 4.14 plots the actual drain to source current for a range of gate to source voltages and drain to source voltages. When the gate to source voltage is small, no current flows. That is the reason that the lowest curve on this graph has a 2V gate to source voltage. At lower voltages the transistor is off, and no current flows through the device. As the gate to source voltage increases, the on resistance of the transistor decreases (the slope of the curve increases). For low drain to source voltages, the curve looks linear, like a resistor and matches our model well. However for high drain to source voltages the curve bend over and look more like a current source than a resistor. To analyze circuits that deal with transistors in this operating region requires a more complex model. Fortunately we don't need that more complex model to analyze logic gates, which we look at next.

input	output
0	1
1	0

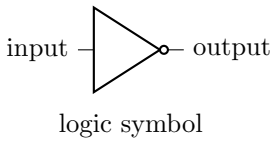


Figure 4.15: Truth table and schematic of an inverter.

4.5 Building CMOS Logic Gates

Now that we know how CMOS transistors work, we can use them to build logic gates. We will start with the simplest gate possible, and inverter, also known as a `NOT` gate. This gate inverts the value of its input. If the input is true, its output is false and vice versa. Figure 4.15 gives the truth table and the logical symbol for an inverter.

It turns out that this logical function is possible, and even easy to construct using MOS transistors. We want the output to be connected to Gnd when the input is a '1' (at Vdd). This is exactly what an nMOS transistor with its source connected to Gnd does. We also need to have the output connected to Vdd when the input is a '0' (at Gnd). Again this is exactly what a pMOS transistor does if its source is connected to Vdd. Thus the CMOS implementation of an inverter uses a pMOS and an nMOS in a simple, symmetric fashion and is shown in Figure 4.16.

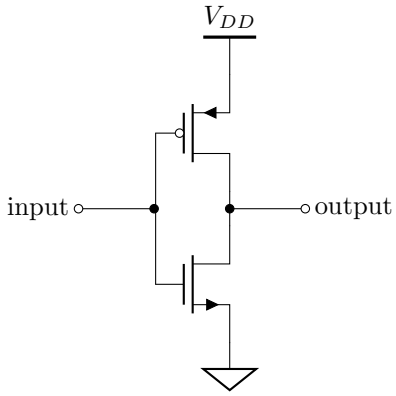


Figure 4.16: MOS Transistor implementation of an inverter.

The inverter example shows how MOS transistors can be used to create a simple logic gate. In fact one can use them to create much more complicated logic gates, and these gates can be wired together to create more complex logical functions. A useful skill is to be able to look at a MOS transistor schematic, and figure out the logical function that those transistor perform, or whether this circuit isn't a valid logical function. To do that we need to understand what constraints a good logical gate should follow. The rules are really pretty simple. Since the output of every logical function is either a 0 or 1 for each input combination, all CMOS logic gates need to follow two rules:

1. The output should always be connected to either Vdd (a value of '1') or Gnd (a value of '0').

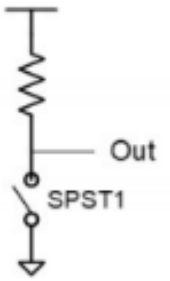
2. The output should never be connected to both Vdd and Gnd at the same time.

This means any CMOS gate needs a collection of pMOS transistors that connect the output to Vdd when the output should be 1 and a set of nMOS transistors that connect the output to Gnd when we want the output to be 0. We also need to make sure that the output is always driven, and it is never driven to both Vdd and Gnd. Notice that in our inverter satisfies these rules. For each input case, at least one of the transistors is on and they are never both on.

If you violate these rules then one of two bad situations happen. If there is an input condition where the output is not connected to Vdd or Gnd then the output node is left floating unconnected to any other node. Since the node has no resistive paths to any supply, any node voltage is possible (there are no device current equations to constrain it), so we can't say whether the output will be 0 or 1, and it might even be in a state that is illegal (say at a voltage that is Vdd/2). If you put Vdd/2 into an inverter, both transistors can turn on, and that inverter will get hot. Transistors will also get hot if you create a gate where some input combination creates a path from the output to both Vdd and Gnd. Now we have a resistive path through the MOS transistors between Vdd and Gnd, and current will flow through these transistors. Now the output voltage will depend on the resistance ratio between these paths, and again the output won't be good digital values.

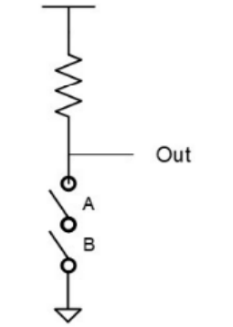
If you want to figure out the logical function of a CMOS gate, create a truth table that includes all the input combinations. For each input combination check to see if the output is connected to either Gnd (enter 0 as the output) or Vdd (enter 1 as the output). Once you have completed all the inputs you will have the truth table of your logic gate. If there cases where the output isn't driven, or both paths are on at the same time, it is not a valid gate.

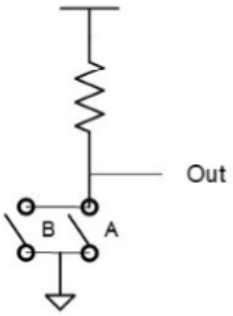
Problem 4.4 When is the output a logical 1?
Determine whether the switch should be connected or disconnected to set the output to logical 1.



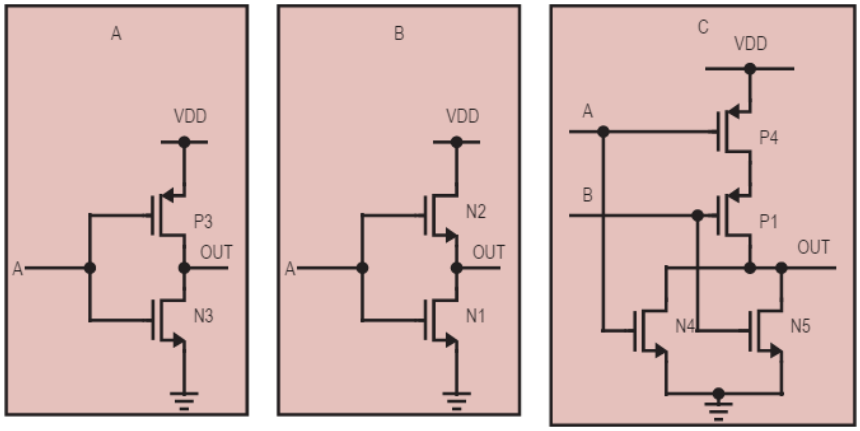
Problem 4.5 When is the output a logical 0?
For each of the following two circuits, determine when the output is a logical 0. Choose one of the following options:

A. When A is connected.
B. When B is connected.
C. When A and B are connected.
D. When A or B is connected.
E. When neither are connected.

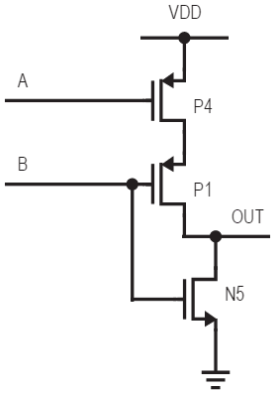




Problem 4.6 Which are valid logic gates?
By using the rules described above, determine which of the following three CMOS circuits are valid logic gates.



Problem 4.7 More practice: Is the following a valid logic gate?



4.5.1 How to create a logic gate?

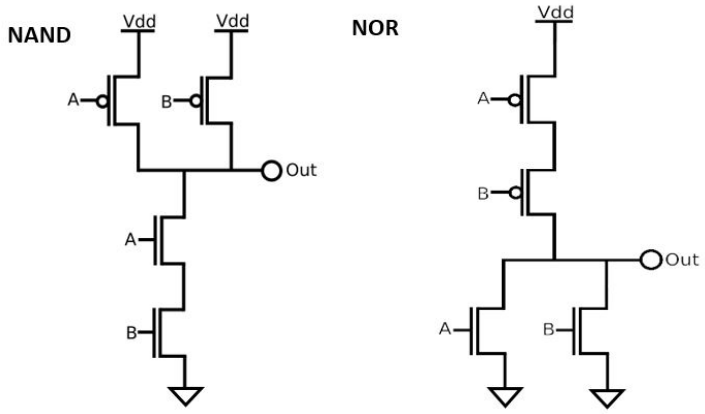
It is not that complicated as long as you follow a few simple rules. First notice that nMOS transistors connect the output to Gnd and turn on when the inputs are HIGH, and pMOS transistors connect the output to Vdd when their input is LOW. This means that 1 inputs can only cause the output to be 0, and 0 inputs can only cause the output to be 1. Said a different way, all MOS gates will invert. They can only make inverters, or NAND gates or NOR gates. To make an AND gate, you need to make a NAND gate and then add an inverter to its output!

When building a logic circuit, its helpful to ask two questions: What conditions should cause

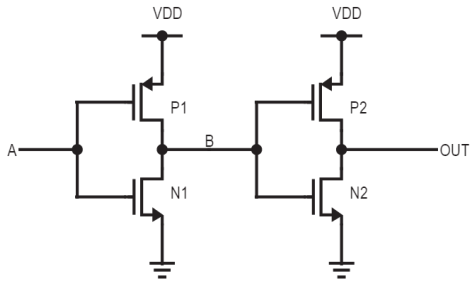
the output to be 0? We use the answer to this question to create the nMOS transistor circuit that connect the output to Gnd. If two inputs both must be true for the output to low, then we need to create a circuit that only connects the output to Gnd when both transistors are on. This condition occurs when the two transistors are connected in series. Then there is a path to from the output to Gnd only when both transistors are on. If the output should be low if either of the inputs are on, then we should create a circuit where the two transistors are put in parallel, where the source of both transistors goes to Gnd, and the drain of both transistors connects to the output. In this case when either transistor is on, the output will be connected to Gnd.

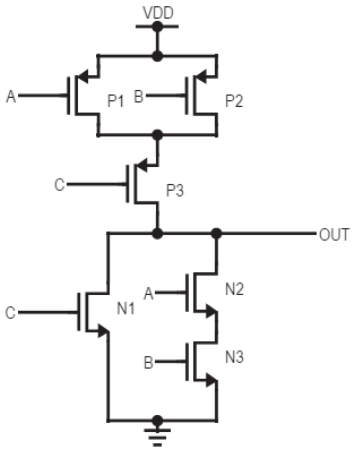
What conditions should cause the output to be 1? We use the answer to create the pMOS transistor circuit that connects the output to Vdd. If two inputs both must be LOW for the output to HIGH, then we need to create a circuit that only connects the outputs when both transistors are on. This condition occurs when the two transistors are connected in series. Then there is a path to from the output to Vdd only when both transistors are on. If the output should be HIGH if either of the inputs are LOW, then we should create a circuit where the two transistors are put in parallel, where the source of both transistors goes to Vdd, and the drain of both transistors connects to the output. In this case when either transistor is on, the output will be connected to Vdd.

Using these rules gives the following transistor circuits for a NAND and NOR gate.



Problem 4.8 Write out the truth table for the following circuits. Then, write a logical expression relating the inputs and the outputs.





4.6 Solutions to practice problems

Solution 4.1:

The light in this circuit is on. Here, the positive terminal of the battery is connected to the gate of the nMOS transistor, so the voltage at the nMOS gate is 5V. The source of the nMOS transistor is connected to ground. Therefore, the gate-to-source voltage v_{GS} of the nMOS transistor is 5V - 0V = 5V. Knowing that the threshold voltage V_{th} of an nMOS is typically between 0.4V and 1V, we can see that $v_{GS} > V_{th}$ and therefore the nMOS is on. Since it is on, there is a path from the left terminal of the lightbulb to ground through the nMOS. The right terminal of the lightbulb is connected to 5V. Therefore, there is a voltage difference across the terminals of the lightbulb and the bulb is on.

Solution 4.2:

For the first circuit, the resistance from A to Gnd is infinity. This is because the bottom-left nMOS transistor in the circuit must be off since its gate voltage is equal to its source voltage (so v_{gs} is 0). Since this bottom nMOS is off, it can be modeled as an open switch and so the resistance through it is infinity. However, the resistance from B to Gnd is $2R_{on}$. This is because both nMOS transistors have $v_{GS} > V_{th}$ and so they are both on. The path from B to Gnd through these two resistors consists of the R_{on} of both transistors in series.

For the second set of circuits, the resistance from A to Gnd is $2R_{on}$. The bottom left nMOS is on because its gate is at a logical 1 while its source is connected to ground, and therefore its v_{GS} is positive. Since the Ron of this transistor is presumably small, we also then know that the source of the transistor above it is at a voltage close to ground, and the v_{GS} of that transistor is also positive. Therefore both transistors are on and the path from A to Gnd through these transistors is the series combination of each of their R_{on}. The resistance from B to Gnd is infinity, because both transistors have $v_{GS} = 0$ and are therefore off.

For all the circuits, recall that there is always infinite resistance between the gate and the source and infinite resistance between the gate and the drain.

Solution 4.3:
The resistance from A to Vdd is infinity because the top PMOS is off.

The resistance from B to Vdd is $2R_{on}$ because both PMOS transistors are on.

Solution 4.4:

The output is set to logical 1 when it is at Vdd. Therefore, the switch should be disconnected. (Since there is no current flowing through the resistor in this configuration, there will be no voltage drop across it and the output will be at the same voltage as Vdd.)

Solution 4.5:

For the first circuit, the output is a logical 0 when A AND B are connected.
For the second circuit, the output is a logical 0 when A OR B is connected.

Solution 4.6:

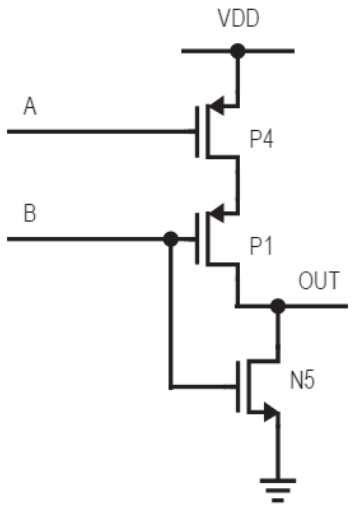
Circuit A is a valid logic gate. When A = Vdd, only the NMOS is turned on, conncting the output to Gnd. When A = Gnd, only the PMOS is turned on, connecting the output to Vdd. For valid inputs ('1' or '0'), the output is never connected to both Vdd and Gnd at the same time. Hence, it fulfulls both requirements for a logic gate. (This circuit is the commonly used inverter.)

Circuit B is not a valid logic gate. When A = Vdd, the output is connected to both Vdd and Gnd at the same time (both the NMOS turn on). This breaks the fundamental rule of logic gates (and in fact could also break your circuit if you actually tried to build it.)

Circuit C is a valid logic gate.

To check this, you need to consider all combinations of valid inputs for A and B and determine which gates are turned on for each case in order to find what the output is. The table below summarises the output for all possible input combinations. The output is a valid value for all of them, and at no time connected to both Vdd and Gnd, hence it is a valid logic gate.

A	B	OUT
1 (Vdd)	1 (Vdd)	0 (Gnd)
1 (Vdd)	0 (Gnd)	0 (Gnd)
0 (Gnd)	1 (Vdd)	0 (Gnd)
0 (Gnd)	0 (Gnd)	1 (Vdd)



Solution 4.7:
It is not a valid logic gate.

We can check by using the same method as before, testing the output for all valid input combinations, as summarized in the table below. We find that for one of the combinations, the output is left floating and connected to neither Vdd nor Gnd. Hence, it is not a valid gate.

You might realise that this problem is fixed in circuit C of the previoous problem. For every PMOS on the upper half pulling the output up to Vdd, there must be a corresponding NMOS on the bottom pulling the output to Gnd in the correct way to ensure this state does not occur.

A	B	OUT
1 (Vdd)	1 (Vdd)	0 (Gnd)
1 (Vdd)	0 (Gnd)	Floating! (not connected to Vdd or Gnd)
0 (Gnd)	1 (Vdd)	0 (Gnd)
0 (Gnd)	0 (Gnd)	1 (Vdd)

Solution 4.8:

1.)
The truth table is shown below. The logical expression is OUT = (A!)! = A.

A	B	OUT
1	0	1
0	1	0

2.)

The truth table is shown below. The logical expression is $OUT = (C \;||\; A \;\&\&\; B)!$. There are many equivalent expressions, but you should always be able to write out an equivalent truth table for whatever you come up with.

A	B	C	OUT
1	1	1	0
1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	0
0	1	0	1
0	0	1	0
0	0	0	1

Chapter 5

Numbers, Computers, and Coding

Chapter 4 just showed that the voltage output of some switch circuits is either Vdd or Gnd, and that it is often useful to think about these outputs as a Boolean variable—it is either True (Vdd) or False (Gnd), or 1(Vdd) or 0(Gnd). It then introduced a new type of device, a MOS transistor: an electrically controlled switch. Using MOS transistors enables us to build circuits that perform any function on Boolean variables. But who wants to do logical functions of Boolean variables? It turns out that everyone does, since combining these functions allows you to build computers, networking equipment, and all the portable electronics we know and love. This magic is possible because we group multiple Boolean values (where each Boolean value is called a *bit*) together and use this collection of Boolean values, or bits, to represent a number, color of the screen, character, etc.. In fact a big part of Electrical Engineering is thinking about how to best represent different types of information in bits for electrical transfer and storage. This area of study is called *codes*, and the study of the limit of codes is called Information Theory.

This chapter starts by exploring different ways a device built from MOS transistors can represent the value "3," or "1,000,000." It then has a brief description of how a computer works, but breaking everything down to logic operations, and how technology scaling made computers cheap. Then we come back to representing numbers, and tackle the problem of representing negative numbers, like "-31," and some of the interesting things that can happen when you represent integers with a fixed number of bits. The end of chapter will look at other types of codes, focusing on the types of codes you need to use to control your LED display.

In electronic devices (devices with MOS transistors on them), the series of symbols that we use to send messages are going to be in the form of **bits**. Since a single bit can only be either be 0 or 1, it doesn't really communicate much, so a collection of bits will be necessary to send more complicated information. The next few sections look at different possible codes, each code is a different trade-off between robustness, flexibility, privacy, and compactness.

5.1 Codes For Representing Numbers

This section presents two codes that are used to represent a number in a computer. The first is a *Unary Code*, and while it is very simple and useful, it requires a large number of bits to represent a large range of numbers. The other, a *Binary Code* is much more compact, and is widely used to represent numbers inside of computing machines.

5.1.1 Unary Code

One of the simplest codes is a *Unary Code*. In this code there is a bit for each value you want to represent. For example, to represent the numbers 0-3 would require 4 bits as shown in Table 5.1. This representation is sometimes also called a *one-hot code*, since in many situations only one bit would be one at any time.

Number	Unary Code
0	0001
1	0010
2	0100
3	1000

Table 5.1: Unary Code for a 4-bit System

The advantage of this system is that it's easy to use the code directly to control individual hardware. If there were four LEDs in a display that you want to control, and you want the bottom LED on for 0, and the top LED on for 3, it is easy to use each bit to control the state of one LED. This type of code can also be used to allow multiple LEDs to be on at one time. Notice if a Unary Code is use to represent a value, it is possible to encode multiple values at the same time. In the display example, this four bit code could indicate that the the '3' and '0' LEDs should both be on (1001). We will use this type of representation later to control the LEDs in our LED display.

Unary Codes are often used when building hardware that needs to go through a time sequence of steps. In this context each bit indicates which step should be execute at this time. As you will see in Section 5.6 we will need to use this type of time sequence control for the LED display as well.

This code is very flexible (it can represent multiple outputs at once) and easy to interpret (each bit means another number). In its pure unary form it is also somewhat error tolerant. If only a single value should be high, any bit error (flip) will either cause two or zero values to be high, which can be detected. However, it is not compact (the number of bits is linear in the number of items you want to represent) or private (everyone who can see the bits knows the value being transmitted). The next section talks about binary numbers, which fix the compactness problem, and a later section will look at how to add error correction to the system.

5.1.2 Binary Numbers

Binary coding is a more efficient way of storing integer values and is very similar to how humans represent numbers. A majority of the people in the world use a **decimal** or **base-10** numbering system. This means that we represent each digit in the number can have potentially 10 different values, which are 0-9, and multiple digits are used to represent a large number. Each digit has a different weight. To get the actual final quantity of a number, for example 145, we start on the right most digit of the number which is 5, multiply by 10^0 (which is just 1). Then we move left to the next digit 4, multiply by 10^1, then add it to 5, which we get 45. Then we move left again to the digit 1, multiply by 10^2, then add it again to our accumulating sum, which we finally get 145.

$$145 = 10^2 * 1 + 10^1 * 4 + 10^0 * 5$$

We can technically continue to do this as we continue to add more and more digits, and the more digits we have, the larger the number we can represent.

10^2	10^1	10^0
1	4	5

Table 5.2: Decimal 145

Since each bit can only represent two values, 0 and 1, we want to do the same thing but now in base 2, not base 10. We can use the same procedure we used before, but now the column values will increase by 2x not 10x: instead of multiplying each digit value by 10 to the power of the digit index, we multiply each bit value by 2 to the power of the bit index. For example, let's take the 8-bit value 00101010. To convert this value into a base-10 number that makes sense to humans, we can do the same process in base-2 to determine the actual value of this binary value. Remember, we start at the right most bit. Then as we shift left, we multiply that value by 2 to the power of the number of bits we've shifted from the right most bit, then accumulate that sum.

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
0	0	1	0	1	0	1	0

Table 5.3: Binary 00101010

$$42 = 2^7 * 0 + 2^6 * 0 + 2^5 * 1 + 2^4 * 0 + 2^3 * 1 + 2^2 * 0 + 2^1 * 1 + 2^0 * 0$$

This is great because in unary coding, 8-bits can only represent 8 numbers, but with binary, the largest number that can be represented is 11111111, which is $2^7 * 1 + 2^6 * 1 + 2^5 * 1 + 2^4 * 1 + 2^3 * 1 + 2^2 * 1 + 2^1 * 1 + 2^0 * 1$, or 255. While the number of bits in a unary code is linear on N, the number of symbols the code needs to represent, a binary code only needs $log_2 N$ bits. Representing a thousand values requires 10 bits, not a thousand.

There are many ways to convert a decimal number to binary. The easiest is to use a calculator/computer that does the conversion for you. If you have to do it yourself, there are some simple iterative procedures you can use for the conversion. One starts by determining the left-most bit, or most significant bit (MSB) first, and the other starts by finding the right most bit, or the least significant bit (LSB) first. Both work by subtracting off the value of one bit, and then looking at the remaining value to generate the rest of the bits. Let's use the number 23 and convert it to an 8-bit binary number as an example.

The left most bit first process starts by comparing the number to the left most place value, which is 2 to the power of that index value, which will be 2^7 in our example. If the number is larger than this place value, put a 1 in this column, subtract this value from the number from the current number. If the current number is smaller than the place value, put a 0 in the column, and pass the current value to the next column. This basically reduces the number by the value represented by this bit. The result is then passed to the next bit position where the step is repeated. In our example, the value is 0 and the remainder is 23. This procedure repeats, next at the MSB-1 column, until all columns have values.

The other option to convert numbers is similar, but generates the number from the right most bit or LSB first. Each iteration looks at whether the number is odd. If it is odd, it puts a one in that column, subtracts one from the number, and then divides the number by 2. If the number is even, it puts a zero in the column, and then divides the number by 2. This step removes the value

or the LSB, and then shifts the resulting number one position to the right so the next column is now the LSB. The divided number now represents the bits of the number is the LSB was removed. Repeating this procedure will produce all the bits of the number as shown by Table 5.4.

Left Bit First Conversion.

Comparison	Output	Next Value
$23\text{-}2^7$	0	23
$23\text{-}2^6$	0	23
$23\text{-}2^5$	0	23
$23\text{-}2^4$	1	7
$7\text{-}2^3$	0	7
$7\text{-}2^2$	1	3
$3\text{-}2^1$	1	1
$1\text{-}2^0$	1	0

Right Bit First Conversion.

Odd?	Output	Next Value
23	1	11
11	1	5
5	1	2
2	0	1
1	1	0
0	0	0
0	0	0
0	0	0

Table 5.4: Conversion from Decimal 23 to 00010111

Problem 5.1 : Representing numbers in binary

1. Find the 4-bit binary representation of 15.
2. Find the 8-bit binary representation of 15.
3. Find the 8-bit representation of 101.

5.1.3 Binary Arithmetic

Addition and subtraction in binary are very similar to their decimal counterparts. You add by carrying 1s and subtract by borrowing 2s (instead of borrowing 10s).

For example, let's examination how addition works when adding 11 and 3, which are 1011 and 0011 in binary respectively:

Step 1

1	0	1	1
0	0	1	1

Step 2

		1	
1	0	1	1
0	0	1	1
			0

Step 3

	1	1	
1	0	1	1
0	0	1	1
		1	0

Step 4

	1	1	
1	0	1	1
0	0	1	1
	1	1	0

Step 5

	1	1	
1	0	1	1
0	0	1	1
1	1	1	0

Table 5.5: Adding Binary Numbers

We get the final result 1110, which is converted to 14 in decimal. This checks out because 11+3=14.

And, for subtraction, we will subtract 3 from 9:

Step 1

1	0	0	1
0	0	1	1

Step 2

1	0	0	1
0	0	1	1
			0

Step 3

	2		
~~1~~ 0	0	0	1
0	0	1	1
			0

Step 4

	~~2~~ 1	2	
~~1~~ 0	0	0	1
0	0	1	1
			0

Step 5

	~~2~~ 1	2	
~~1~~ 0	0	0	1
0	0	1	1
		1	0

Step 6

	~~2~~ 1	2	
~~1~~ 0	0	0	1
0	0	1	1
	1	1	0

Step 7

	~~2~~ 1	2	
~~1~~ 0	0	0	1
0	0	1	1
0	1	1	0

Table 5.6: Subtracting Binary Numbers

In the end, we get the binary value 0110, which is 6 in binary. This is correct because 9-3=6. People almost never do binary arithmetic. Its power is in the fact that the operations that are needed to perform the operation can be expressed as Boolean functions, which means that they can be implemented by MOS transistors. An adder can be built by building the same hardware for each column, since we did the same steps to generate each bit of the sum.

This logic per column needs to have three one bit inputs. Obviously it need the bits from the two numbers being added together, but is also needs to know if the previous (lower significance) column generated a carry. Let's call the operand A and B, and the carry from the previous carry, CarryIn. The column logic needs to generate two outputs, the sum for this column, Sum, and whether it will generate a carry to the next column, CarryOut. It is not hard to create the logic that generates Sum and CarryOut. If you are interested in the logic, just create the truth tables for Sum and CarryOut as a function of the three inputs.

The important result is that it is easy to build adders out of logic gates, and it is easy to build FSM controllers from logic gates. It is also easy to build memory from MOS transistors, in fact probably all the memory you use today is made from MOS transistors, unless you still have some spinning disks. Both non-volatile memory (FLASH) and both types of memory (DRAM and SRAM) are built from MOS transistors. This means that all the parts needed to create a computer, network, cellphone, tablet, etc. can be built from MOS transistors. This coupled with the fact that the technology used to make MOS transistors continues to improve, has made computing and other electronics cheaper and cheaper. If you are interested in learning more about how to use MOS transistors to build computers, you might want to take EE108.

Problem 5.2 : Binary arithmetic

Compute the following binary arithmetic problems. You can check your answer by converting your binary solution into decimals.

1. 0010 + 0001 = ?
2. 00001111 + 01100101 = ?

5.2 Computing and Technology Scaling

It is hard to believe, but it was not that long ago (in the early 1970s) when electronic computing devices were very rare and expensive. Accountants who needed to add a large number of numbers used mechanical adding machines, and students were taught how to multiply numbers using a nomagraph called a slide rule. A slide rule was a primitive analog computer that allows someone to estimate the answer to multiplication, division, and even trigonometric problems. But the invention of a planar integrated circuit in the 1960's soon changed all that.

Rarely does a single technology takes over the world in quite the way that integrated circuits (IC) have done. From their invention around 60 years ago, ICs are now found in almost everything we do, from cards to computers, and will become even more prevalent in the future. This section will briefly look at why this happened, and the tremendous opportunities it creates. It has made it possible to create reasonable computers with wireless connections that sell for under one dollar, and made it possible for people like us to "make" impressive, hardware/software/mechanical systems. Our current challenge is to figure out how to transform the capability of this technology into something of value for the consumer without becoming lost in complexity.

To understand why integrated circuits have taken over the world, we will break this question into two pieces. We will first explore why electrical or electronic solutions beat the other alternatives, and then why integrated circuits are the preferred way to build electronic systems. The first question is easy to answer: electrical signals are easier to move/control than controlling items with real "mass". The easiest way to show this is to compare a mechanical adding machine to an electronic calculator. But perhaps a better example is to compare a mechanical computer to an electronic one.

Charles Baggage conceived of a mechanical device that could compute astronomical table of numbers in the early 1810s, and built a small prototype. Even with substantial funding from the government, he was not able to complete the machine. In the 1840's he created an improved design, Difference Engine II, which was never built. In 2000, the Museum of Science, London commissioned the construction of this machine, which turned out to function properly.[1] This mechanical computer is shown in Figure 5.1. While it is impressive that this machine worked, and the spiral carry chains visible in the back of the machine are beautiful, it ran at a few operations per sec, and weighed a few tons.

By the 1940s people started trying to build electronic computers. They used the vacuum tubes, which were the first electrically controlled switching device that was invented. The ENIAC was an example of this machine, and it also was large. It used 17,000 tubes, 70,000 resistors, 10,000 capacitors, and 6,000 mechanical switches. This large machine could add around 5,000 numbers in a second. While the first computers were built using vacuum tubes in the late 1940s, they quickly migrated over to using transistors after the transistor was invented in 1947. Vacuum tubes have a

[1] https://en.wikipedia.org/wiki/Difference_engine

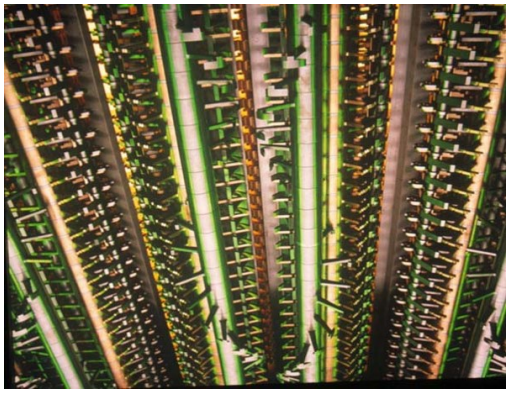


Figure 5.1: Photograph of the Difference Engine II. The calculating section of Difference Engine No. 2 has 4,000 moving parts (excluding the printing mechanism) and weighs 2.6 tons. It is seven feet high, eleven feet long and eighteen inches in depth.

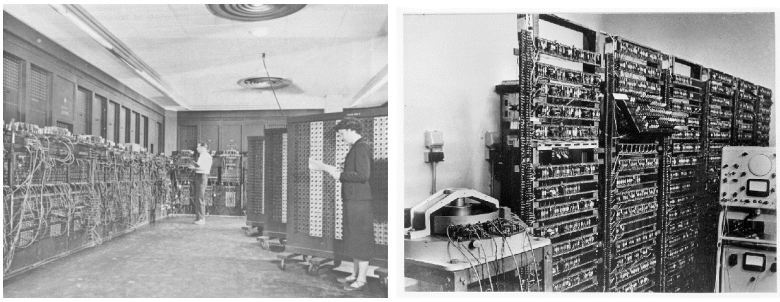


Figure 5.2: Picture of the ENIAC (left) and the Manchester Transistor 1 (right)

hot filament, like a light bulb, and like incandescent light bulbs they burn out. This makes building a large computing machine difficult, since inevitably one of the many tubes would burn out and the machine would fail. Using transistors greatly improved the reliability of the electronic switches, and makes it possible to build larger, more complex machines. Around 6 years after the transistor was invented, working computers based on transistors were built. Not only were these transistor based computers much more reliable than the tube based machines, they were also much smaller, and dissipated many orders of magnitude less power (100W-1KW, vs 160kW for the ENIAC).

Using electronic circuits to generate the results rather than mechanical linkages made the resulting systems much faster, lighter, and more reliable (once we moved away from tubes). As a result mechanical computing devices became obsolete. As electronic components, like transistors became more mature, their cost dropped, but there always was a minimum cost of handling each component that was needed for a machine. This minimum cost of each component meant that the total system cost was proportional to its complexity.

The integrated circuit removed this coupling between complexity and cost. Invented in 1958 by Jack Kilby at Texas Instruments (a leading manufacturer of transistors at the time), and made practical in 1961 by Bob Noyce at Fairchild Semiconductor, an integrated circuit enabled one

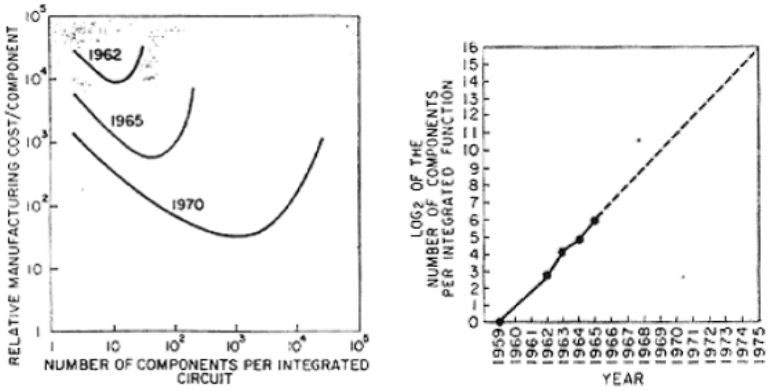


Figure 5.3: Figures from Gordon Moore's paper in Electronics which predicted that the number of transistors one should put on an integrated circuit which would minimize their cost would increase exponentially over time. The original paper claimed that the number would double every year. Later this was changed to a doubling every 1.5-2 years, but this exponential scaling of the number of transistors has continues into the late 2010s, 50 years since the paper was published.

to create an entire circuit on one piece of silicon. Furthermore, since the circuit was created through a printing process, like printing a color picture, the cost of the circuit depended on the area it required, and not the number of components it contained: multiple devices could be created (printed) in parallel, so the cost of the device didn't really depend on its complexity any more. As the resolution of the printing process improved over time, individual transistors became smaller and cheaper, making the current generation computing systems cheaper, and enabling to the creation of more powerful, and more complex computing systems to be built.

In 1965, less than five years after a practical method of creating integrated circuits was invented, Gordon Moore, one of the founders of the integrated circuit start-up company called Intel, published a paper predicting that the cost per logical operation of an integrated circuit would decrease exponentially in the future.[2] His data, shown in Figure 5.3 is now referred to as Moore's Law.

It is this relentless improvement in integrated circuit technology that has created the information ecosystem that we enjoy today. It is the technology that powers all our computing and communication devices. It is also the technology that makes possible single chip computing platforms like the Arduino, that we use in this class.

5.3 Arduino

"Arduino" is the combination of a software development environment, a series of microcontroller boards, and a library which provides a common set of useful functions across all of the hardware boards. The fact that the design and development package are open-source, combined with its emphasis on a easy-to-use programming model has helped make Arduino the standard for hobbyist electronics projects in recent years. While there are now Intel and ARM powered Arduino designs, most Arduino boards still use Atmel microcontrollers.

[2]Electronics, Volume 38, Number 8, April 19, 1965

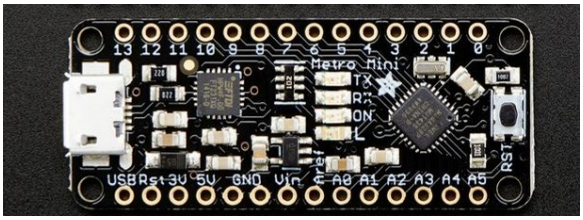


Figure 5.4: A picture of a MetroMini, the version of Arduino you will use in the class

Your lab kit includes a couple Arduino Metro Minis, which contains everything you need to make an embedded computer-controlled system - an Atmel micro-controller, power supply and regulation, a USB connection for power and programming, and input/output pins which make it easy to connect to both digital and analog circuits.

The Metro Mini is designed to be very compact, and fits nicely into your breadboards. Larger Arduino boards include a standard set of header pins positioned precisely to be "plug and play" compatible with a set of accessories called Shields. Through shields, you can add everything from wireless networking to displays to your Arduino. If you are interested in Arduino Shields, you should check them out on the web.

You can program your Arduino using your laptop or desktop computer through its USB port. Development for Arduino is all done through the Arduino IDE, discussed at:
http://arduino.cc/en/Guide/Environment.

5.3.1 Arduino IDE

The Arduino IDE has been crucial in the Arduino's widespread success. It is designed to allow people with little to no programming or hardware experience quickly get software up and running. It has created a set of library routines that make driving values on pins of the chip and reading values from other pins of the chip relatively easy. They also have libraries to communicate results back to your computer, and other stuff that makes building computer-controlled hardware easier. While it is a popular programming environment, it doesn't have the normal debugging environment that you might be used to. Since the microcontrollers that it runs on and the programs are both relatively simple, it provides only a simple programming/debugging environment.

To get started with your Arduino, you will need to download the Arduino development environment from http://arduino.cc/en/Main/Software and install it. Once you have downloaded the software, you need to configure it for our board. The metro mini was designed by adafruit to be compatible to the Arduino Uno. So please run the IDE and configure it for an Uno: Click "Tools → Board" and select "Arduino Uno". The other thing you might need to set is the serial port used to communicate with your Arduino. This again is under the "Tools" tab, and if you are lucky, there will only be one port listed. Note that the port name can change (or dissappear entirely) depending on which USB port you plug the Arduino into, and the IDE won't automatically correct it. Worse, it will sometimes appear to download code correctly, even when you have the wrong port selected (this occurred most frequently on Macs). If unplugging, re-plugging, and selecting the right port doesn't work, you may have to reboot.

To check to see if you have set everything up correctly, open the "Blink" program example. "File→Examples→01.Basics→Blink. You should be able to change the delay in the program and

change the blinking rate of the LED on the board.

The Arduino programming language is a slightly restricted dialect of C++. If you're not familiar with C++, that's fine - it has a very similar syntax to Java. One advantage of using Arduino is that there is a lot of code out on the web for Arduinos, so when you are trying to do something, there often is code that is close to what you are trying to do that you can use and modify. In this class we encourage you to start with a program that does something related to what you need, and then modify it, to do exactly what you want. All we require is you acknowledge the source of code, and that you don't just copy a friend's code. Under the "File" tab, you will find the "Examples" dropdown menu. In that section you will find many programs that do interesting tasks. We encourage you to look at these examples, or search the website to better understand the programming language and some of the special functions that Arduino provides.

Thorough documentation for all of the built-in functions is on the Arduino reference page: http://arduino.cc/en/Reference/HomePage.

When you start a new program, the Arduino IDE provides two functions `setup()` and `loop()`. The `setup()` function is called once when the Arduino boots up (during power-up or when reset), and `loop()` gets called repeatedly after that. Generally, `setup()` will contain initialization code, and `loop()` will contain some code that reads inputs and responds by setting some outputs.

Arduino systems communicate with the outside world using an *input/output pin*, or *I/O pin*. This pin is the interface between a microcontroller and another electrical circuit. It can be configured in the microcontroller's software to be either an input or an output. On the Arduino, this configuration is accomplished using the `pinMode()` function. The following subsections describes three functions: `pinMode()`, `digitalWrite()` and `digitalRead()`. The details of all of these can be found in the Arduino reference at https://www.arduino.cc/en/Reference/HomePage. (If you're reading a printed copy of this document, the links all go there.)

5.3.2 Looking Behind the Curtain (optional reading)

When you click "Verify", the IDE combines the code you've written with a template `main.cpp`, which you can find under the installation directory as `hardware/arduino/cores/arduino/main.cpp`.

The meat of `main` is just:

```
setup();

for (;;) {
    loop();
    if (serialEventRun) serialEventRun(); // Handle serial port I/O
}
```

It passes this to the AVR-GCC compiler, which is Atmel's version of GCC for AVR microcontrollers.

If you are familiar with C++, you might want to know that most standard C++ syntax is supported, including classes, references, etc. However, some important things are not:

- There is no heap memory allocation. We've only got 2.5kB of RAM!
- There is no implementation for the C++ standard library (i.e., `iostream`, `string`, etc.)
- Exceptions are not supported.

If you want all of the details, look at Atmel's documentation for the AVR-GCC compiler: http://www.atmel.com/webdoc/AVRLibcReferenceManual/index.html

5.3.3 Output pins

When configured as an *output pin*, a pin provides V_{DD} or 0 V to whatever is connected to it (if anything). When the pin provides V_{DD}, we say that it is in the `HIGH` state. When the pin provides 0 V, we say that it is in the `LOW` state. We set the state of an output pin using the `digitalWrite()` function.

An output pin achieves this by making a connection to V_{DD} or a connection to ground internally, inside the microcontroller. This, in turn, is accomplished with transistors whose drains are connected to the output pin, as shown in Figure 5.5.

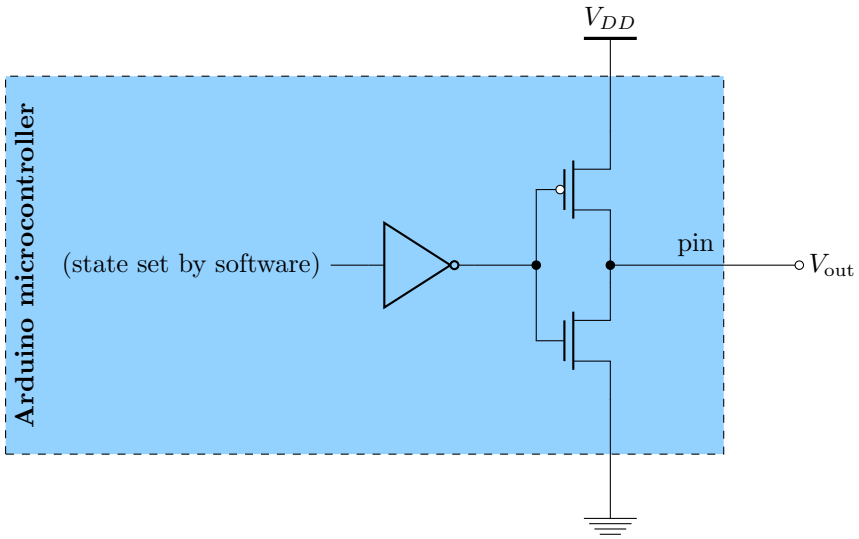


Figure 5.5: Schematic of an I/O pin when configured as an output

When the output is set to `HIGH`, the PMOS transistor turns on and provides a connection to V_{DD}. When the output is set to `LOW`, the NMOS transistor turns on and provides a connection to ground.

Output resistance

If these transistors were ideal, then they would provide a direct connection to V_{DD} or ground, depending on which output state (`HIGH` or `LOW`) was set by software. Of course, these transistors aren't ideal; as we understood from our discussion of transistors, they have some (hopefully small) *on resistance*. Furthermore, the outputs of microcontrollers are typically designed to power low loads, so this resistance isn't negligible.

The resistance thus "seen" by a device connected to a pin is called the *output resistance* of the pin. Of course, there is not just one output resistance—it depends on the state of the pin. When the output is `HIGH`, it is the resistance to V_{DD} that is relevant ($R_{\texttt{HIGH}}$ in Figure 5.6); in our model,

this is the on resistance of the PMOS transistor. When the output is `LOW`, it is the resistance to ground; in our model, the on resistance of the NMOS transistor. Often these nMOS and pMOS transistors are sized so their resistances are similar, which is the case with our chip.

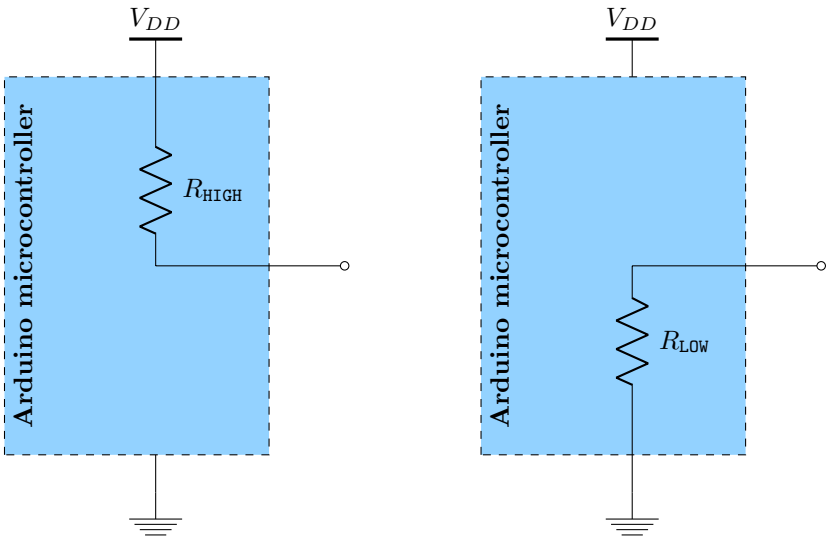


Figure 5.6: Equivalent circuits showing the output resistance when the output is high (left) and low (right)

Measuring the output resistance

Say we wished to measure the resistance of an output pin when it is set to `HIGH`. First, we would need to set the output pin to high, using the `pinMode()` and `digitalWrite` functions. Then, we can measure the resistance with an ohmmeter, as shown in Figure 5.7.

Recall, however, that what we're measuring isn't a resistor, but a transistor that is on. Therefore, it is polarized, and unlike with a resistor, the polarity of the leads matters. The ohmmeter will try to push a test current from the red lead to the black lead, so we should place our leads so that this test current goes in the expected direction. When measuring the resistance of a PMOS transistor, this means putting the red lead at the source (VDD), and the black lead at the drain (the pin). Similarly, when measuring the resistance of the NMOS transistor, the red lead should be at the drain (pin), and the black lead at the source (ground).

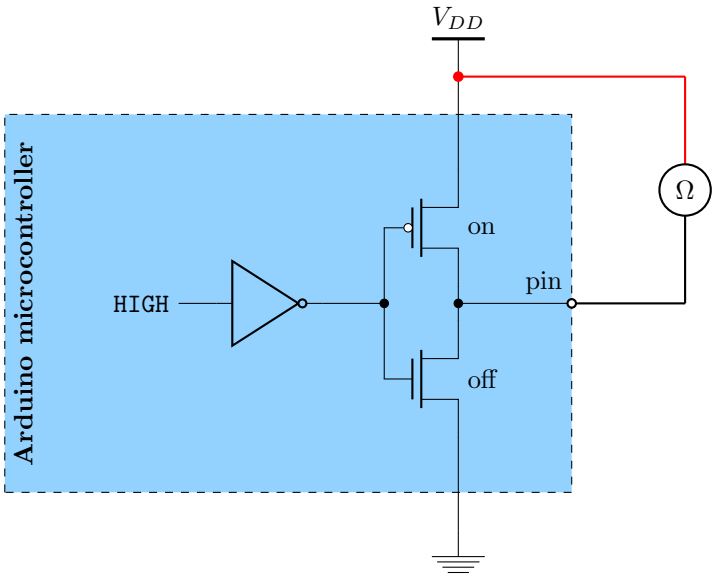


Figure 5.7: Measuring the output resistance of a pin when it is set to `HIGH`

For many devices, including driving a single LEDs, this output resistance isn't too much of an issue: we need a external resistance to limit the current in our diode, and that external resistor is larger than the pin resistance. However there are devices which require large current, like a motor. For the useless box, the motor requires 60mA when it is running, but can require up to more than .4A if it is stalled. The pins of our Arduino can't drive this large current, and we need to place a driver circuit in between the output pin and the motor, so that the motor can draw enough current. Such a driver circuit often uses *power transistors*, transistors that are designed to provide higher current (lower drain-source resistance) than those in the microcontroller.

5.3.4 Input pins

An *input pin* reads the voltage on the pin as if it were a voltmeter, and returns either `HIGH` (`1`) in software if the voltage is close to V_{DD}, or `LOW` (`0`) if it is close to 0 V. An input pin can be read using the `digitalRead()` function.

Note that the value returned by `digitalRead()` is only well-defined when the input pin voltage is close to V_{DD} or 0 V. The precise meaning of "close" varies between microcontrollers, but for the Adafruit Metro Mini[3] as it's used in our circuit, the input pin voltage needs to be at least $0.6V_{DD}$ to qualify as `HIGH`, and at most $0.3V_{DD}$ to qualify as `LOW`. In the middle (say, at $0.45V_{DD}$), the behavior of the pin is undefined.

An (ideal) input pin takes (approximately) no current, like the gate of a transistor or a voltmeter. In the projects we do in this class, modeling the input pin as ideal is a good approximation.[4]

[3] More precisely, it's the ATmega328, which is the microcontroller used in the Adafruit Metro Mini. The full datasheet can be found at http://www.atmel.com/images/Atmel-8271-8-bit-AVR-Microcontroller-ATmega48A-48PA-88A-88PA-168A-168PA-328-328P_datasheet_Complete.pdf; this information is on page 313.

[4] The input leakage current is specified in the datasheet to be at most 1 µA.

Connecting a switch to an input pin (pull-up resistors)

How do we connect a switch to an Arduino? This is not as obvious as it sounds: remember that switches govern connections, not voltages. So simply connecting the switch between the pin and some other point in the circuit, say, ground, would *not* work (Figure 5.8). When the switch is closed, the input pin would be shorted to ground, and V_{in} would be 0 V, which is fine. However, when the switch is open, the input pin is floating, and we have no idea what its voltage is. (In practice, it will "remember" the last voltage it was driven to, which in this case is 0 V.)

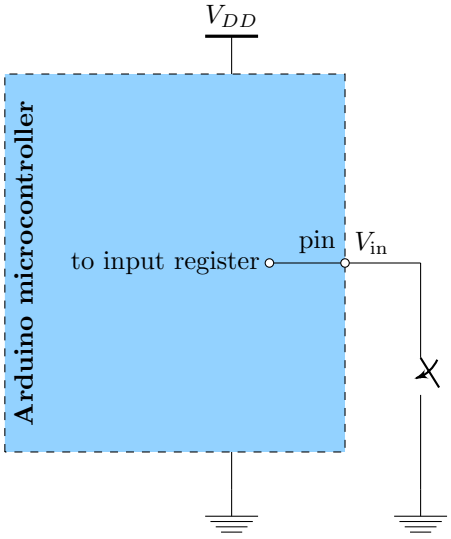


Figure 5.8: A switch circuit that does not work

Consider, instead, the circuit shown in Figure 5.9. When the switch is closed, the pin is still shorted to ground, so $V_{\text{in}} = 0\,\text{V}$. However, this time, when the switch is open, no current flows through the resistor R_{pu} (there's nowhere for it to go, since the input pin takes no current), so the voltage drop across the resistor is zero, so $V_{\text{in}} = V_{DD}$. Thus, we can reliably distinguish between the closed and open states of the switch: when the switch is closed, the input pin will read `LOW`, and when the switch is open, it will read `HIGH`.

You can think of the role of the resistor as being to enforce a "default" state on the pin. When we directly connect the pin to ground, V_{in} is set to 0 V by that connection—the resistor won't get in the way (current will flow through it, but that doesn't affect the input pin voltage). In the *absence* of such a connection, though, the resistor comes into play, *pulling up* the voltage V_{in} to V_{DD}. For this reason, the resistor R_{pu} is often referred to as a **pull-up resistor**.

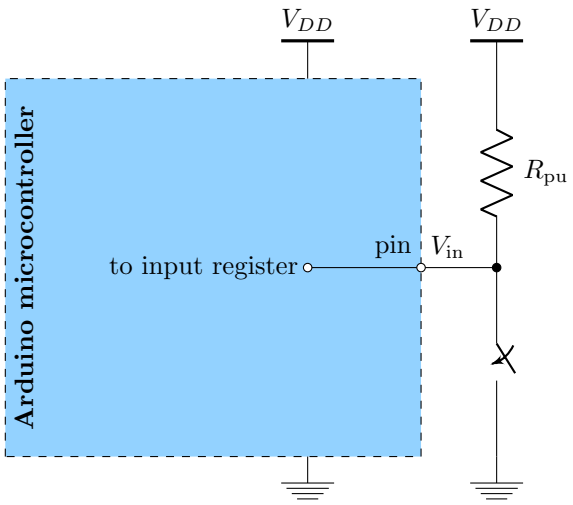


Figure 5.9: Using an external pull-up resistor

We could also do this the other way round, connecting the switch between V_{DD} and the pin, and the resistor between the pin and ground, so that would be a "pull-down resistor". However, connecting the switch to ground and the resistor to V_{DD} tends to be the more common practice. It's so common, in fact, that many microcontrollers (including the Adafruit Metro Mini) implement a pull-up resistor internally, inside the microcontroller, as shown in Figure 5.10. The pull-up resistor can be enabled or disabled in software. To enable it, we pass the `INPUT_PULLUP` constant as the second argument to `pinMode()`: `pinMode(pin, INPUT_PULLUP)`. To disable it, we pass the constant `INPUT` instead: `pinMode(pin, INPUT)`.

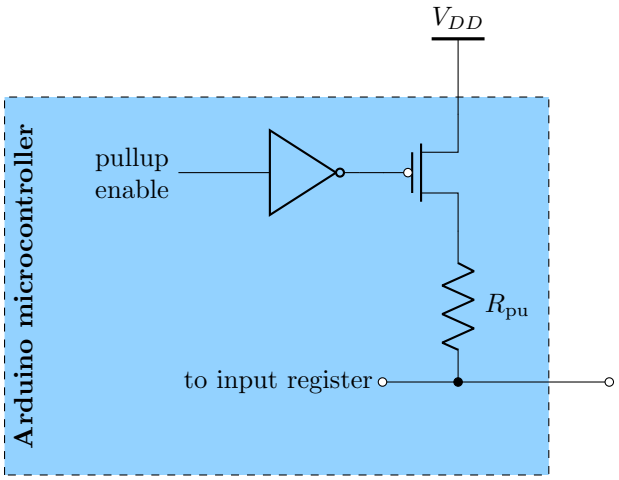


Figure 5.10: Internal pull-up resistor schematic

The enabling/disabling of the internal pull-up resistor is implemented using—you guessed it—

a PMOS transistor, connected to VDD. When the transistor is off, it looks like an open circuit, effectively removing the pull-up resistor from the circuit.

If you use the internal pull-up resistor, you naturally don't need the external one. This makes it acceptable to connect your switch as in Figure 5.8, so long as you enable the pull-up resistor, making the equivalent circuit in Figure 5.11.

Finally, a couple of cautionary notes. First, we haven't yet talked about the value of the resistance R_{pu}. This is partly because it doesn't really matter—for any reasonable resistor value, this circuit will function as we described; the precise resistance only affects how much current flows when the switch is closed. But there is another aspect to this: the internal pull-up resistor isn't specified to have a precise value. Typically, it's on the order of tens of kiloohms; for the Adafruit Metro Mini, it's specified to be between 20 kΩ and 50 kΩ. So you shouldn't use the internal pull-up resistor if you need a precisely-known resistance.

Secondly, the pull-up resistor obviously creates a path to V_{DD}, which can sometimes make a pin you intended to be an output pin seem like it's working even when you forgot to include `pinMode(pin, OUTPUT)` in your `setup()` function. This is because, in a slight quirk of the Arduino, when a pin is configured as an *input* pin, the `digitalWrite()` function actually enables (`HIGH`) or disables (`LOW`) the pull-up resistor. (That is, `pinMode(pin, INPUT_PULLUP)` is actually just shorthand for `pinMode(pin, INPUT); digitalWrite(pin, HIGH)`.)

So if you forget to configure a pin to output, you're actually just enabling and disabling the pull-up resistor. For a multitude of reasons this is a *bad idea*: First, as discussed above, the pull-up resistor does not have a precisely-defined resistance. Also, when the pull-up resistor is disabled, the pin is floating—which is acceptable for an LED, but if you're using this pin to drive the gate of a power transistor, it will mean you have a floating gate. For this reason, it's important always to remember to configure the `pinMode()` of all pins that you're using in your code.[5]

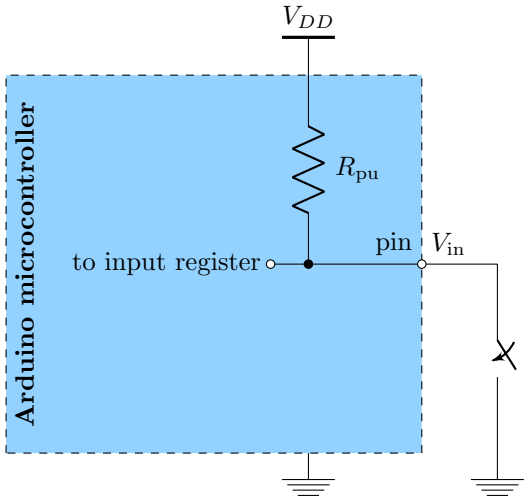


Figure 5.11: Using an internal pull-up resistor

[5]If you don't configure a pin, it defaults to being an input. However, to make your code easier for others to read, you should explicitly configure your input pins, too.

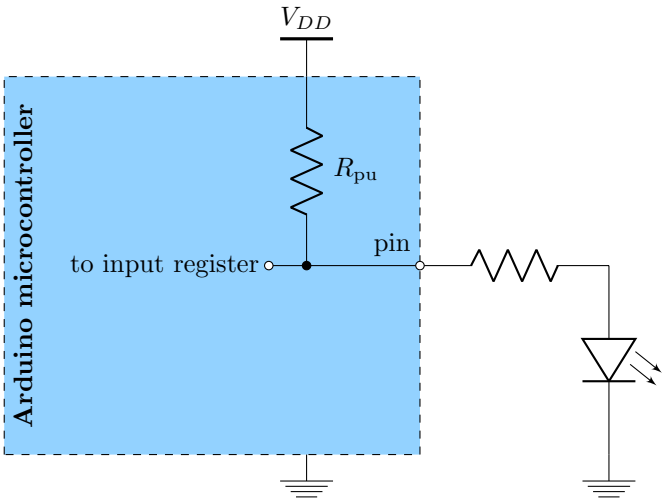


Figure 5.12: When you forget to define `pinMode(pin, OUTPUT)` (bad, don't do this), turning the pin on and off with `digitalWrite()` will just enable and disable the pull-up resistor.

5.4 Binary Numbers, Revisited

While the earlier section provided an overview of how numbers are represented in a computer in a binary form, it left out two important details: how negative numbers are represented, and the errors that are possible when a finite number of bits are used to represent a number. Most modern computers use 32 bits or 64 bits to represent an integer internally. These representations are usually large enough that errors from the bit representation are rare (but they still occur). Unfortunately the Arduino only uses 16bits to represent most numbers by default, so these types of "overflow" errors are more common. This section looks at both of these issues, to help you recognize the "funny" errors they can cause in your Arduino programs.

5.4.1 Integer Overflow

Since all computers use a fixed number of bits to represent a number, it is possible that the number you want to represent is larger than the set of numbers that can be represented by these bits. For example, the largest 4 digit decimal number is 9999. Analogously, the largest 4 bit binary number is $1111 = 2^0 + 2^1 + 2^2 + 2^3 = 1 + 2 + 4 + 8 = 15$. The number of bits (places) that a computer has to store a number depends on the computer. Common numbers of bits computers use include 8, 16, 32, and 64. For example, Arduino MicroMinis use 16 bits to store integers. It turns out that the maximum value that can be stored in an n bit binary number is equal to $2^n - 1$ (this can be shown by using the formula for a sum of a geometric sequence). Thus, the largest unsigned integer that the Arduino can store is $2^{16} - 1 = 65535$.

This limited number of places to store numbers can cause problems when we want to get numbers larger than the maximum or smaller than the minimum.

For example, when we add 1 to the maximum integer, we get **integer overflow**.

1	1	1	
1	1	1	1
0	0	0	1
0	0	0	0

Table 5.7: Integer overflow. $15 + 1 = 0$?

This problem stems from the fact that the leftmost value place has a carried 1 that can't be added. Similarly, when we subtract 1 from the minimum integer, we get **integer underflow**.

~~2~~ 1	~~2~~ 1	~~2~~ 1	2
0	0	0	0
0	0	0	1
1	1	1	1

Table 5.8: Integer underflow. $0 - 1 = 15$?

This problem stems from the fact that the leftmost value place has to borrow a 2 that doesn't actually exist. What happens in these cases is the addition is done *modulo* 2^n where n is the number of bits used. In modulo arithmetic, all factors of 2^n are factored out. $2^n + 3 = 3$ and $52^n = 0$. So it is possible to add to a large number and get a small number, or subtract from a small number and get a large number. While this seems bad, and it isn't good, there are some interesting rules that still hold. For example assume that the variable in question is a timer, and is measuring the amount of time the processor has been on. You are using that number to figure out how long it has been since you last looked at the timer by subtracting the old time from the current timer. This operation will always return the correct value of time, as long as the time interval (time between reading) is less than the total number of time steps of the timer. It is a property of modular arithmetic that interval computations work, even when overflows occur!

If you don't believe it, try it out yourself. Assume that your last measurement was 1110, and the current measurement is 0001. When you subtract 1110 from 0001, you will get 0011, which is the three time units that passed between those measurements.

Problem 5.3 : More binary sums

1. 1111 + 0101 = ?
2. 11110001 + 00001111 = ?

5.4.2 Negative Numbers

The binary number representation presented in Section 5.1.2 only represents positive numbers. In general most people want to represent both positive and negative integers. One solution is to dedicate the most significant bit (left most bit) to represent the sign bit of the number. A positive number would set the sign bit to 0, and a negative number would set that bit to 1. For example, the number 3 in 8-bit can be represented as 00000011. If we wanted to represent -3, we flip the sign bit, giving us 10000011. This type of representation is called sign-magnitude representation,

and it is used in some number systems (see real numbers later in this section). Yet it is not the representation used for integers. The reason is simple. In sign magnitude representation, we need to look at the sign bit before we can add the two numbers together, since we need to know whether we need to setup an add, or subtraction. We could make the add operations faster if we chose a representation of negative numbers where this control operation is not needed. The representation of this form is called two complement numbers, and is used to represent negative integers.

Two's complement numbers takes advantage of the type of overflow that occurs in modular arithmetic. Instead of needing to "subtract" negative numbers during addition, it uses the representation that occurs from the subtraction to represent the negative number. An example will make this more clear. Suppose we want to represent -1. To generate the bits we will use, let's state with 0 and subtract 1 from it. As shown in the previous subsection, this yields 111 in a 4 bit representation, or 15. While this might seem like a weird way to represent -1, If you add 15 to any number, n, the results is going to be n-1, exactly what we want. 2+15 mod 16 = 1. For an 8 bit number, -1 would map to 11111111 or 255, for a 16bit number it would map to 1111111111111111 or 65535.

Since numbers wrap around due to integer overflow and underflow, we can just designate that certain numbers are negative, and use modulo (or for 4 bit numbers, subtracting 16) to determine negative numbers! In other words, if the leading bit is 1, then we know we are in negative number land and the number can be calculated by subtracting $2^{\text{number of bits}}$ from the number. This concept of using this circular wrap-around for negative numbers is known as **two's complement**.

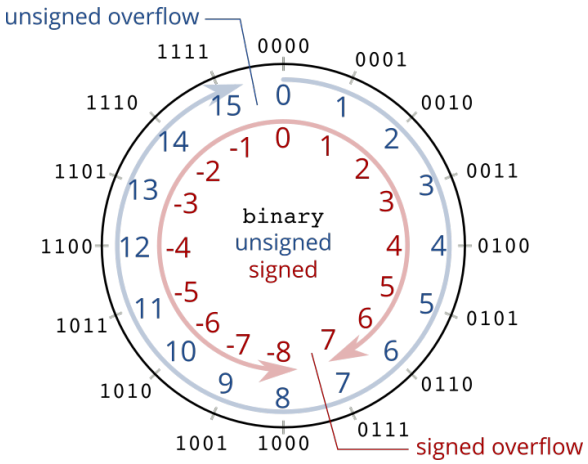


Figure 5.13: Modular arithmetic creates a circle of numbers, and maps multiple numbers to the same bit representation. As a result the bit string 1111 can be considered either 15, or -1 in two's complement notation.

Generating the two-complement version of a number is actually pretty simple. Given the number you want to complement, say 4, or 0100, we first complement all the bits, yielding 1011. This is called the one's complement of the number. But this isn't exactly what we want, since if we add this to the original number we get 1111, and not zero. But we can easily get zero by adding one to the one's complement version, since 1111 +1 = 0 in modular arithmetic. So we add one to the one's complement to get 1100, which is -4 in Figure 5.13.

Binary Code	Unsigned Value	Signed Value
0000	0	0
0001	1	1
0010	2	2
0011	3	3
0100	4	4
0101	5	5
0110	6	6
0111	7	7
1000	8	-8
1001	9	-7
1010	10	-6
1011	11	-5
1100	12	-4
1101	13	-3
1110	14	-2
1111	15	-1

Table 5.9: Binary Code for a 4-bit System, both Signed and Unsigned

This mapping of multiple values onto the same string of bits is not good for computers, since often a program wants to know if one value is larger or smaller than another. Interpreting the value as 15 or -1 will yield different answers to this question. Since we don't want the bits interpreted to be both -2 and 14 at the same time, we're going to define two conventions for interpreting bits: unsigned and signed.

Unsigned convention is exactly what you've seen in the previous section (binary without any negative numbers). Overflow and underflow occur at the maximum integer and 0. This sets all bit combinations as positive integers.

Signed convention is a little bit different. We want a way to represent negative numbers, so all numbers whose most significant bit (the leftmost bit) is 1 will be negative. This also means that overflow and underflow occur at a different location (see the two's complement circle). With signed numbers when adding two positive numbers yields an overflow, a negative number results, and vice versa.

Notice that arithmetic overflow can still exist in a two's complement numbers. The largest positive number is 0 followed by only 1's, so for an 8-bit system, it would be 01111111. If you add 1 to that value, you will get 1000000, which is a negative number in two's complement. This is overflowing past the upper bound.

Furthermore, the largest negative number is 1 followed by only 0's, so for an 8-bit system, it would be 10000000. If we add negative one to this value (00000001 is 1, inverting gives us 11111110, adding one gives us 11111111), will give us 01111111 where we lose an extra 1 due to overflow. This an example of underflowing past the lower bound.

Be very cautious when performing arithmetic in 16 bit integers because it's possible to go past the upper and lower bounds. The maximum number of positive values expressible by a fixed number

of bits in a two's complement system compared to an unsigned integer system has halved because those values are now being used to represent negative numbers. If you know all your numbers are going to positive, you should use unsigned integers (unit) in your C code.

For example, consider the interpretation of binary 1011. Using an unsigned convention we get $2^0 + 2^1 + 2^3 = \mathbf{11}$. Using a signed convention we notice that the most significant bit is 1. Thus, we subtract $2^{\text{number of bits}} = 2^4 = 16$, from 11, so we get $11 - 16 = -\mathbf{5}$. The procedure is the same for binary 0111, but since the most significant bit is zero, the result from both procedures yields the same answer, $2^0 + 2^1 + 2^2 = \mathbf{7}$. It should be easy to see that in a two's complement representation, the most positive number is a 0 followed by all 1s, while the most negative number is a 1 followed by all 0s.

One additional point that is worth mentioning is that the same procedure used to convert positive numbers to negative numbers works in reverse as well. It will also take negative numbers and make them positive. Just invert each bit and add one to the result:

1. Decimal 5 = Binary 0101
2. Invert bits: 1010
3. Add 1: 1011
4. Decimal -5 = Binary 1011
5. Invert bits: 0100
6. Add 1: 0101
7. Decimal 5 = Binary 0101

Problem 5.4 :

1. Represent the following numbers in 4-bit two's complement form, then in 8-bit two's complement form.
   - 5
   - -5
   - 8
   - -8
2. Represent the following numbers in 8-bit two's complement form.
   - 15
   - -15

The great advantage with two's complement is that arithmetic is much easier. No matter the sign, the inputs are simply added together. If you are not convinced try a few examples on the modular arithmetic wheel in Figure 5.13. When two numbers need to be subtracted, we simply

perform the operation to turn the number we wish to subtract into a negative number, and then add the numbers together.

Using a 4-bit system, let's first compute 2 - 6 in binary. We should get -4 at the end, but let's go through the process in binary.

We establish first that 2 is 0010 (2^1), and 6 is 0110 ($2^2 + 2^1 = 6$). Next we need to convert 6 into -6, and we do this by inverting the bits: 1001 and adding one to get 1010. Finally, we can perform the addition of 0010 and 1010, which we should get 1100.

What is 1100 in decimal? Let's convert this back into a positive number and see. First we invert the bits 0011 and add one which gives 0100. 0100 is 4 (2^2), so our answer is -4, which agrees with the decimal arithmetic solution and shows how two's complement addition indeed works.

5.4.3 Real numbers

In addition to representing integers, programs also deal with real numbers. These are the numbers that can express an infinite number of values between 0 and 1, as well as the large values that integers represent. This large range of values that real numbers need to express is a challenge for its representation system, especially if one wants to represent these numbers in a limited number of bits. Many standards have been created which define these representations. All these representation use the same basic approach of creating a *floating point* representation. Here the bits are broken into three groups. One group is the sign bit,S, so floating point number are generally sign-magnitude representation. The next set of bits are the mantissa bits, M. Like the bits in an integer, they determine the value of the number, but the value of the mantissa is forced to lie between 1 and 2. This seems quite limiting until you consider the last set of bits, which are the exponent bits, E. The final value is $-1^S \cdot M \cdot 2^E$. Since the exponent, E, can be positive or negative, the representation can express numbers that are much larger than 1, or much less than one.

5.5 Error Correcting Codes

After our deep dive into binary representation of numbers, we are going to pop up and think about other characteristics you might want from a code you use. For communication, being able to reliably communicate information between machines is very important. When sending digital information through a communication channel, all packets of data will be transferred and received successfully in a perfect world, but this is not always the case in reality. Various reasons such as noise, outside interference, or simple system glitches causes information to be lost at some probability. This is a complex problem, since noise or system issues will affect the packets differently and the system will need to be able to determine if any packets have errors, etc. When trying to solve a complicated problem it is sometimes helpful to create a example with many of the same essential issues as the original problem, but is cleaner and simpler to think about. Often times the solution to this simplified problem can then be extended to the real problem that one wants to solve. So rather than trying to attack the real problem of communicating in a noisy world, we will look at a simplified version of this problem which is communicating over an **erasure channels**. An erasure channel is a very nice bad channel. It does lose blocks of data we send over the channel, but it lets us know which blocks of data it corrupted. As a result the receiver knows which blocks of data, which we will sometimes call packets, are good, and which are bad. The blocks of data are also numbered, so the receiver also knows which data blocks they received error free. With a K erasure channel, if one sends N blocks of data through that channel, the receiver might only get N-K blocks of data.

While the receiver will get N-K good blocks of data, no one knows which blocks of data will be dropped. As a result, in order to guarantee that the data will make it through, it seems like (and is correct) that we need to transmit every data block at least K+1 times to guarantee it makes it to the receiver. That way, the worst-case scenario would be if the K packets of data that are lost are all copies of the same piece of data, leaving one copy left. Therefore, if we can send 12 packets of data at once and we can potentially lose 2 pieces of data, we need to replicate all data at least 3 times, meaning we can only send 12/3 = 4 packets of data. Each block of data would be copied three times. While this works, it seems not optimal: since we duplicate all the data, most of the data being received will be copies of data that has already been received. But this does provide a lower bound on the amount of data blocks that can be transmitted = $\frac{N}{K+1}$. Perhaps there is a better way of sending data?

We can easily determine the upper bound of data that can be transmitted through this link is (N-K), or the number of packets that are successfully sent through the communication channel, since we can't get more bits through the channel than the bits we receive. However, in order for the receiver to get this number of bits, each packet of data that gets through the channel must have unique information in it. This need for unique information is hard to square with the requirement that you need to send each data block multiple times over the channel.

This seems like an impossible problem, until you realize that there is no reason you can't send multiple pieces of data in the same data packet. So rather than sending one data block of information each time, you can send multiple data blocks by adding them together.[6] Now we can make sure each data is replicated in multiple packets, while at the same time ensuring that each packet is unique (up to N-K).

Let's do an example of this:

Say that we need to send the numbers X1, X2, and X3 through a communication channel that sends 5 data words but can lose any two words. If we send the following 5 packets, it's possible to recover X1,X2 and X3 no matter what 2 packets we lose, since every piece of data is transmitted three times:

- X1
- X2
- X3
- X1+X2+X3
- X1+2*X2+3*X3

For example, if we only received the values, X1, X2, and X1+X2+X3, we can recover X3 by subtracting X1 and X2 from the final value.

What if we only received X1+2*X2+3*X3, X1+X2+X3, and X3? We can simply perform (X1+2*X2+3*X3) - (X1+X2+X3) - 2*(X3) to get X2. Then using X3 and X2, we can get X1 from the second value.

Bottom line is that as long as we have 3 linearly independent equations with 3 unknowns, we can easily determine the three values.

[6] If you are paying close attention you will realize that adding packets together might use more bits than before, since it could generate bigger numbers. When this is done in practice, the addition is done modulo arithmetic so the numbers don't really get larger

This shows that if we can send N packets and we lose at most K packets, we can acquire N-K=M packets. Another way of thinking about how to solve this problem is by using linear algebra to create our new values that are a function of the other pieces of data. If X is our input data and Y is the data that we send, we can perform Y = AX where A is a NxM matrix.

$$Y = AX$$

$$\begin{bmatrix} x1 \\ x2 \\ x3 \\ x1+x2+x3 \\ x1+2*x2+3*x3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix}$$

Finally, to convert the original values back when we lose data, we take the rows from A of the data that we recovered, invert the new matrix, then multiply it by Y.

$$X = A^{*-1}Y^*$$

Where * refers to the data you've received. If you lose the data x2 and x3, the procedure you perform will be:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}^{-1} \begin{bmatrix} x1 \\ x1+x2+x3 \\ x1+2*x2+3*x3 \end{bmatrix} = \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix}$$

The are many other coding tricks for error correction, including ways of coding numbers so it is possible to detect and correct single bit, or even double bit errors. Like the example show above, while at first it would seem that this protection would require duplicating every bit (so it can be recovered), these duplications can be "shared" in a surprising small number of additional bits. For more information about coding, please see classes like EE178.

5.6 Time Multiplexed Codes

Imagine you have an LED screen that has 1920 x 1080 resolution, meaning that is the number of pixels in the X and Y dimensions. Each pixel has a red-green-blue (RGB) value, so has 3 values for each pixel, and each value will require 8 bits. How do we go about controlling each pixel in the screen?

The first approach would be to wire each individual pixel and color to a control circuitry, but this is not a very scalable solution. For the 1920 x 1080 monitor, the number of pins needed to power the screen will be 1920 * 1080 * 3 * 8 or around 48 million pins. The maximum number of pins you will be able to attach to the glass screen is in the thousands, so this just isn't going to work. We are going to need a different approach to this problem that uses less wires. Yet at the same time, we still need to send information about all the values of all the pixels to the display.

There are many tricks that are used to "solve" this problem. The main one is to not send the data all at once, since we don't have the number of wires needed. If the wiring between the computer and display is limited, we will need to transmit many values (bits) one one wire. The common approach to accomplish this is to send data sequentially in time on the wire. This concept is called *Time Multiplexing*, or **Serial Communication**. Many common communication protocols

take advantage of this concept, such as Ethernet, USB, I2C, SPI, JTAG, and many more. The number of wires needed to transmit data is relatively small because the data sent through those communication channels is sent as packets over time. In fact the connection between your computer and its display is probably HDMI or Display Port today. If you cut these cables, you would find a small number of wires which communicate all the display values from the computer to the display. These serial communication protocols generally have an integrated circuit driving one side of the cable, and another integrated circuit on the other side of the cable. These integrated circuits use complicated logic to serialize and de-serialize the data going over the wire.

But even when you use these chips, eventually one needs to drive the actual display. Since these displays are generally made from glass, and not silicon, the kind of logic you can build in the display is limited. They do manage to put a transistor in each g, r, b pixel on the screen, which is used to store an analog value that represents that pixel's brightness. How those analog values are loaded into the screen is very similar to the method we will use to control out LED display, which doesn't have any transistor associated with it. We will first describe the LED case, and then come back to the LCD screens at the end of this section.

5.6.1 LED Display

Suppose you want to control a 8 x8 LED display. This display has 64 LEDs, which is much larger than the 20 I/O pins our Metromini Arduino has. To solve this problem we clearly need to use time multiplexing to put multiple pieces of data on the same wire. But how can we do that in a way that doesn't require any transistor (or at least not many transistors) in the display? That is the challenge.

To solve this challenge requires a few tricks. The first, and probably most important is to realize that our sensors, like our eyes, are built to work on the time scales that we operate in, which from a computer's perspective is pretty slow. Lights that blink faster than 30 times/sec will appear to be constantly on, which is how displays work and is the reason that movies seem continuous. This slow human response, called optical persistence will allow us to only light up some of the lights at different times, and still have the display seem to be fully on. The other tricks will be to divide time into 8 slots and light only 1/8 of the lights each slot, take advantage of the fact that diodes can conduct current in only one direction, and use unary codes (see Section 5.1.1) to encode the time slots. Let's see how this works.

The main problem we need to solve is how to control N^2 lights with only $2N$ wires. The approach we will take is to break time into N time slots, and during each time slot, only try to light up one row of the display. Since each row of the display only has N LEDs, we can control all these LEDs using only N wires, one for each column in the area. As the time slice changes, we need to turn off the row currently lit, and then turn on the LEDs we want lit in the next row. As long as we get through all the rows fast enough (greater than 60 times a second) everyone will see all the lights in the display on at the same time!

At first this solution seems like it will take less than $2N$ wires, since we need N wires to select which LED is on, but we only need $\log N$ wires to indicate the current time period. For an 8 x 8 display, this would only require 3 wires. This encoding of the time slot is indeed possible, but it would need to be decoded to be able to select which row should be turned on during each time period. To remove the need for this external decoder, we will use a unary code for the time slice, and this will require N wires.

To understand how we can select a single row of LEDs to light up in each time period, lets

examine a row of four LEDs. The positive end of all the diodes are connected together and brought out as the row signal, as shown in Figure 5.14. The negative end of each diode is connected to a wire that runs vertically, which we will connect to at the bottom of the array. So for a single row (1) of four (4) LEDs, we have a 1 x 4 array, with one row connection on the left, and four connections at the bottom of the array. If we wanted to turn on the first LED from the left, we need to allow current to flow through the LED by connecting its column wire low (of course through a current limiting resistor that is not shown) and connect the row wire to Vdd. This will allow current to flow from Vdd through the diode to Gnd. Setting any of the column wires low (through a current limiting resistor) will cause that LED to light up. Driving that column wire high will cause the voltage drop across the diode to be zero, so it will have no current flowing through it, and it will not emit any light.

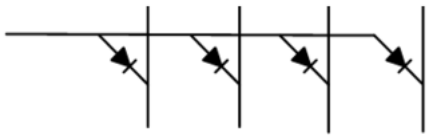


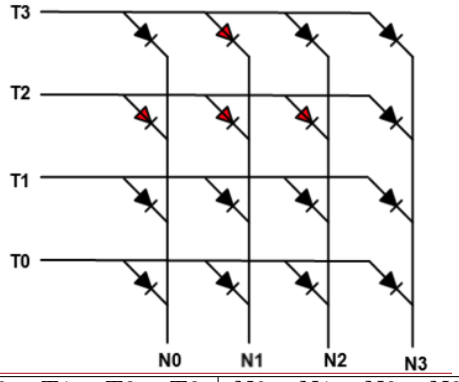
Figure 5.14: One row of LEDs

Now that we understand how to control one row of LEDs, let's look at a 4 x 4 array, which is shown in Figure 5.15. Like before we have four column wires, now labeled N0 - N3, but now we have four row lines, labeled T0-T3. Since we are time division multiplexing, we will only light one row at a time. This means when T0 is high, all the other T lines must be 0 — the definition of a unary code. Notice when a row line is at Gnd, the positive end of all the diodes in the row are at Gnd, which means that the voltage across these diodes will be either 0 or negative. This means that the diodes will be off, just as we wanted.

Assume we want to light up the diodes that are red, and leave the black LEDs off. For this pattern, when T0 is high, we need to drive all the N lines high, since no LEDs in this row should be on. During the next time period when T1 is high (and all the other T lines are low) again all the N lights should be high to keep all the LED off. When T2 is driven high, N0-N2 should now be driven low, to light the three selected LEDs, and finally during T3 only N1 should be driven low, to light the final remaining LED. Notice with this configuration we can control 16 LEDs with 8 pins. Four pins control the rows while the other four control the columns. In this design it is critical that only one row is driven to Vdd at a time, while any column can be driven low to turn on the desired light in that row.[7] If more than one row is driven to Vdd at a time, then we could no longer uniquely light each LED, since setting a column to ground would light the corresponding LED in both rows.

We almost have everything we need to drive our 8 x 8 LED array. It is easy to increase the size of the 4 x 4 array to make it the right size. But now we have an additional problem. We would like our LEDs to be bright, which means we would like to run them at reasonably high currents. The LEDs we use can handle currents of 20-30mA. This current is compatible with the drive of our

[7]Note that we just chose to drive the positive terminal (T lines) sequentially. It is possible to drive the cathode column low one at a time, and then drive any of the rows high to light LEDs in that column, but then we would need to have the resistors placed in series with the rows, and not the columns to limit the current.



T0	T1	T2	T3	N0	N1	N2	N3
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	1	0	1	1	1	0
0	0	0	1	0	1	0	0

Figure 5.15: Four rows of time multiplexed LEDs, with the values that should be driven on the column lines during each time period to light up the red diodes.

Arduino pins. These pins have a series resistance of around 20 Ω and can deliver more than 40mA of current. When using a 5V power supply of the Arduino, this means we need to add around 80 Ω in series with an Arduino output to drive the N0-N7 lines to get 20mA for a 3V V_F diode (blue, green, white), and get 30mA for a 2V V_F diode (red). At each time period, since only one row line will be driven high, only one LED in each column can be on so the maximum current these drivers will see is around 30mA.

Driving the T0-T7 lines is more difficult. While the current when the line is low is zero, for the time when this line is driven high, it might need to supply the current to all 8 diodes. This can be up to 160-240mA which is well above the drive capability of an Arduino pin. So the Arduino can't drive these wires directly. We will need to solve this problem like we did when we needed the Arduino to drive a motor; we will need to use external transistors which can supply more current. We clearly will need to add a pMOS transistor to each T0-T7 line to connect it to Vdd at the right time. A important question to answer is whether we need to add an nMOS transistor to drive the line to Gnd as well. Avoiding the nMOS transistor would make our design much simpler. Figure 5.16 will help us answer this question.

Assume that A is 0, so !A is 1, so the pMOS device driving the top row is off. Since the transistor is off, we don't know what the voltage of the row is, so it at first seems possible that some of the LEDs might be on. In the worst-case some of the column lines would be low, so if the row voltage was high enough, some LED could be on. To answer this question we use charge conservation (KCL) and say that whatever the voltage is, we know that the current flowing out of the node must be equal to the current flowing into the node. It is here that the fact that the lights are diodes is very helpful. We know that current in a diode can flow in only one direction, and in this circuit it means all diode current flows out of the row node. Since the pMOS transistor is off, there is no current flowing into the row node. Since the current flowing into the node is zero, there can't be

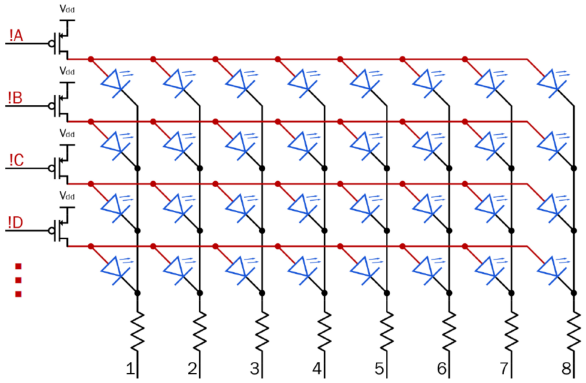


Figure 5.16: Circuitry used to drive the LED display, showing a resistor in series with each column line to limit the LED current, and an pMOS transistor to drive each row line high, and supply the large current that is required. Notice that if the pMOS transistor is off, no external current can flow into the row line.

any current flowing out of the node. This means if the pMOS transistor is off, it is not possible for current to flow through any of the diodes, and no light will come out of that row. We don't need to add a nMOS transistor to make the array work!

5.6.2 Keyboards

Keyboards also use serial communication to connect to your computer. Your keyboard is probably either wireless, or uses a USB (universal serial bus) link to connect to your computer. But we are not going to look at that. Rather we are interested in how someone can build a device with close to 100 nice switches (one for each key) for a few dollars. Here the problem is the opposite of the display problem. In a keyboard, we need to build and read 100 switches and we want to do it very cheaply, which means we don't want to use many wires. The way keyboards solve these problems is amazing engineering. To really understand how a keyboard works, I suggest you take one apart. It is usually not hard to find a discarded keyboard, and they are not that hard to take apart. You generally need to remove a large number of screws in the back of the keyboard, but after that, things get easy.

Since building each switch can be expensive, keyboard don't do that. Instead they have a molded plastic frame, where molded plastic keys are inserted. This combination forms the surface that your fingers interact with, and is shown in Figure 5.17. This plastic forms the frame of the keyboard, but has neither a keyboard feel, or any switches. The feel of a keyboard comes from a sheet of molded elastomer. This sheet provides a little bump under each key. That bump resists the keys motion until enough force is applied when it collapses. This controlled collapse is what give a keyboard it feel.

The actual switches are formed from two plastic sheets that have conductive ink on them, and are shown in Figure 5.18. They are separated from each other by another plastic sheet that has holes at each key position. This middle sheet makes sure that the two sheets with ink on them don't normally touch unless a key is depressed, which physically pushes one sheet down on the other through the hole in the middle sheet. Like integrated circuits, this allows them to build as

Figure 5.17: Picture of the mechanical frame of a keyboard. This is the part that your fingers interact with.

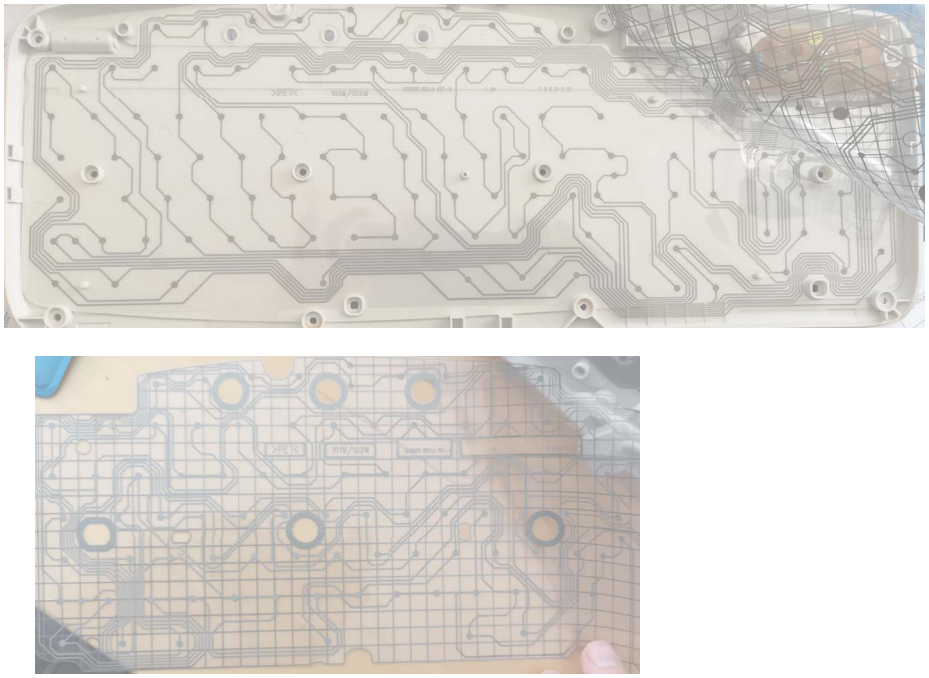


Figure 5.18: Picture of the two plastic sheets that are separated by another sheet with holes in it. The little circles in each of the sheets is where the keys are located.

many switches as they what, for the same cost (they just need to print out the ink pattern on the sheet.

If you look at the conductive lines on the two plastic sheets, you will see that on one sheet the lines run mostly vertically, and in the other sheet they run mostly horizontally: the keyboard is using the same trick we used to reduce wires that we used in our display. By setting one wire low

on the first layer, they can find out if there are any wires that are pulled low on the second layer. If one of those wires is connected to Gnd (the input pins need to have a pull-up resistor to make sure an unconnected wire is pulled up to Vdd) then the key at that crossing position must be depressed. By scanning all the wires in one layer, a microcontroller can determine which key is depressed.

If you were paying close attention to the section on the LED display, you might be wondering whether the fact that a keyboard doesn't have any diode action will mess up our multiplexing scheme. In a keyboard when a key is depressed, it directly connects the column wire to the row wire, where in the LED display, the connection was through a diode. The lack of diodes does restrict how they keyboard works. In the display, is is possible to display all lights "at the same time," which really means we can handle any combination of lights we want to display. The keyboard always works when only one key is depressed, and will work when any two keys are depressed. But there are some situations when three or more keys are depressed, that the computer will not be able to determine which keys are depressed.

Problem 5.5 : Understanding why a keyboard can get confused.

T3	7	8	9	.
T2	4	5	6	/
T1	1	2	3	*
T0	-	+	0	=
	N0	N1	N2	N3

Lets assume we have a simple numeric keyboard which has 16 key arranged in a 4 x 4 grid as shown in the table above, and assume that the switches are connected in a simple grid with the row lines being driven low in different time periods. Like our display, during T0 we will activate the bottom row and connect it to Gnd. We then sense N0-N3 to see if any of those lines are connected to Gnd. If N2 is at Gnd, we know that the '0' switch must be depressed, since that is the only switch that connects N2 to T0.

This arrangement runs into a problem with some three switch configurations. Please figure out what the keyboard would sense if the 4, 5, and 2 key area all depressed at the same time. Remember that the keys are switches, so current can flow in both directions now. Please give what N0-N3 lines are pulled down during each time period.

5.6.3 LCD Displays

Before leaving this chapter, we need to get back to what happens in real computer displays. These systems have millions of pixels, and behind each pixel is a single MOS transistor. The gate of this transistor is connected to a row control line, and there is a line for each row in your screen. The source of this transistor is connected to the column line, and there is a wire for each column in your screen. So like your LED display we control N^2 pixels, with $2N$ lines. Like our display, the row control lines have a unitary code, and raise one line each time period. During this time that analog value for each pixel in that row is driven on the column lines, and since the transistor is now on, this value is driven into the pixel. At the end of this time period, the row line goes low, and all the transistors on that row turn off. The value of each pixel is now stored inside the pixel on a device called a capacitor, which is the subject of our next chapter.

But this does mean that each screen needs to have thousands of wires to connect to it, with circuits that drive both the column and row wires. These wires come from special integrated

circuits that are placed on flexible printed circuit boards that connect to the glass panels. Again it is amazing engineering at work.

5.7 Solutions to Practice Problems

Solution 5.1:
1.)The most significant bit of a 4-bit binary representation has a value of $2^3 = 8$. From then, the next bits have values of $2^2 = 4, 2^1 = 2, 2^0 = 1$ respectively.

$$15 = 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0$$

Therefore, the binary representation of 15 in 4-bit binary is: **1111**.
2.)The 8-bit binary representation of 15 would be **00001111**.
3.)The 8-bit binary representation of 101 is **01100101**.

$$101 = 0 \cdot 2^7 + 1 \cdot 2^6 + 1 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$$

Solution 5.2:
1.) 0010 + 0001 = 0011; In decimal: 2 + 1 = 3
2.) 00001111 + 01100101 = 01110100; In decimal: 15 + 101 = 116

Solution 5.3:
1.) 1111 + 0101 = 10100; In decimal: 15 + 5 = 20
2.) 11110001 + 00001111 = 100000000; In decimal: 241 + 15 = 256

Note that in both cases overflow occurs, and we had to extend the number of bits we are using. In a real computer this often isn't possible (for example, your Arduinos have a limited number of bits you can use for numbers), and we would simply overflow (wrap around). In that case, question one would wrap to 5, and question two would wrap to 1.

Solution 5.4:

1. The 4-bit two's complement form and 8-bit two's complement form of the numbers are as follows:

   - 4-bit: 0101 8-bit: 00000101
   - 4-bit: 1011 8-bit: 11111011
   - 4-bit: You can't! 4-bits is not enough to represent 8 in two's complement - the highest we can go is 7. 8-bit: 00001000
   - 4-bit: 1000 8-bit: 11111000

2. The 8-bit two's complement form of the numbers are as follows:

   - 00001111
   - 11110001

Observe the difference between the positive of a number and the negative of it. Once you've figured out how two's complement works, the rule where you can find the negative version of a number by flipping all the bits and adding 1 will always work (as long as you are still within the allowable range of the number of bits you have!)

Solution 5.5:

In the situation, the keys 4, 5, and 2 are all depressed at the same time. During T0, none of the N0-N3 lines are connected to Gnd as is expected. But during T1, something funny happens. Since key 2 is depressed, N1 is connected to ground as we expect. The problem begins with the fact that key 5 is also depressed. Since N1 is being driven low, the connection at key 5 will drive the row line T2 low (this is the situation that the diode prevents in our LED display). With T2 low, the connection at key 4 will drive N0 low during T1, which is what would happen if key 1 was depressed! During T2 N0 and N1 are both driven low as they should be. At the end of reading all the keys, we know that at least 3 keys were depressed out of 1,2,4,5, but we don't know whether 3 or 4 keys were pushed, and we don't know which three keys were depressed.

Chapter 6

Capacitors

In every device we have studied so far—sources, resistors, diodes and transistors—the relationship between voltage and current depends only on the present values, their previous values were never important. While this is a nice world to work in, and we can solve all of these problems using nodal analysis, it is not an accurate model of the electrical world we operate in. For example, in our model of CMOS inverter, as soon as the voltage on the input goes high, the output immediately goes low[1]: in this model there should be no time delay between the input of the inverter going high, and the output going low. Yet if you measure the input and output voltages of a real CMOS inverter over time, you see that the output doesn't change instantaneously. Instead it changes slowly, and doesn't change until after the input has changed. This delay in the output change, usually called gate delay is an important characteristic of a logic gate, and is what sets the speed of your computer. Clearly we are missing an important component in real electrical circuits.

To capture this component, we need to model how electrical systems store energy. To represent electrical energy storage, we need to add two new types of devices, a capacitor[2] and an inductor[3] This chapter will describe the behavior of a capacitor, and Chapter 9 will describe inductors. Since these elements store energy, their behavior now and in the future depends on past voltage and current. This characteristic enables a wealth of new applications, including estimating the delay of our logic gates, memory, and creating circuits that process musical tones in a frequency-dependent way, allowing us to build an equalizer using just electrical devices. In order to describe this behavior, we'll need to start thinking of voltage and current as *functions of time*, which we might denote $v(t)$ and $i(t)$. After doing this for a while, unfortunately people often become lazy and omit the (t) part, so that v and i are implicitly understood to be functions of time.

6.1 What is a Capacitor?

A capacitor is an device that models the energy that is stored when two wires at different voltages come close to each other, as shown in Figure 6.1. This figure also shows the schematic symbol used to represent a capacitor. Since the two wires don't connect, the element is in some ways an open

[1]The output must now be low since the gate of the nMOS transistor is high, which turns on the transistor and connects the output to Gnd, and the pMOS transistor turns off, so no current flows through the pMOS transistor

[2]Which represents energy storage in electrical fields for those of you with a good physics background.

[3]Which represents energy storage in magnetic fields.

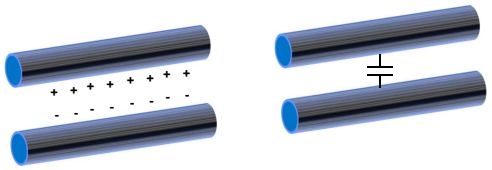


Figure 6.1: A diagram of how capacitance occurs when two wires are close to each other, and how this effect is modeled by a capacitor. When two wires are close to each other, charge is needed to support the potential difference between the two wires. This charge is modeled by a capacitor which is shown in the figure on the right.

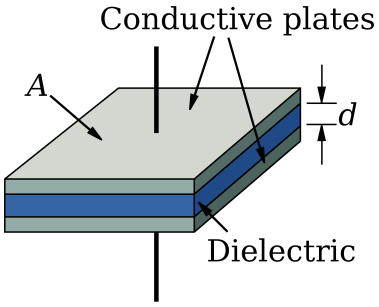


Figure 6.2: Capacitors store energy between two charged plates

circuit: it can't have any current going through it if the voltages on the two wires aren't changing. However to create a potential difference between the two wires requires some + charge to be added on the higher potential wire, and some - charge to be added on the lower potential wire.[4] The amount of charge needed for each additional Volt between the two leads is called the capacitance, $Q = C \cdot V$, which has the dimensions of charge divided by voltage, or Coulomb/Volt which is called a Farad. You should also notice that while a capacitor does store charge in it, the plus and minus charge exactly balance, so the capacitor is still charge neutral.

While all wires and physical elements have some capacitance associated with them, sometimes we want to add an explicit capacitor to our circuit. In that case we want to build a structure which tries to get the most capacitance possible in the smallest space. This is generally accomplished by placing two metal sheets very close to each other, as is described next.

A capacitor circuit element can be built using two parallel plates separated by a dielectric to store energy in an electric field, seen in Figure 6.2.[5] Current can't pass through the dielectric, because it's an electrical insulator. When there is a voltage between the two plates, positive charge builds up on one plate, and negative charge builds up on the other, creating an electric field in the dielectric. The amount of charge Q in coulombs that is stored in the capacitor is related to the voltage across the capacitor v by the equation

$$Q = Cv, \tag{6.1}$$

[4] For those with a physics background, this charge is needed to create the electric field which must exist anytime there is a voltage difference between two conductors.

[5] https://en.wikipedia.org/wiki/Capacitor

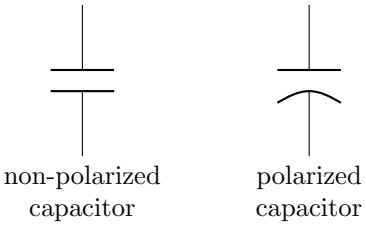


Figure 6.3: Symbols for capacitors

Figure 6.4: Polarized Electrolytic Capacitors

where C is the capacitance of the capacitor. The larger the capacitor, the more charge would be stored for the same voltage.

Geometry of capacitor

The capacitance C of a capacitor is governed by the area of these plates, the distance between them, and the material used for the dielectric, according to the equation

$$C = \frac{\varepsilon A}{d} \tag{6.2}$$

where ε is the dielectric constant, a property of the material, A is the area of the plates, and d is the distance between them. The larger the area and the closer together the plates are to each other, the higher the capacitance.

Capacitors are made of many different materials; you'll encounter several different types in your labs in this class. In most capacitors, there's no positive or negative terminal, just like with resistors.

Polarized capacitors

However, some capacitors are made in a way that means they can only accept voltages in one direction. These capacitors are called *polarized capacitors*. If a capacitor is polarized, we often use a special symbol for it, as shown in Figure 6.3. In particular, in a type of capacitor called *electrolytic capacitors* (Figure 6.4), the dielectric is a very thin oxide layer that would be depleted if a negative voltage is applied, causing the capacitor to start conducting as if it were a short circuit; the resulting huge amount of current then boils the electrolyte, pressure builds up, and the capacitor explodes. In other words, double check the polarity before connecting these!

6.2 Function of a Capacitor

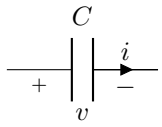
Derivation of $i = C\frac{dv}{dt}$

While a capacitor stores an amount of charge proportional to the voltage across it, we rarely think of capacitors in terms of $Q = Cv$. The reason is that we don't generally think about charge. We generally relate current and voltage. This means the more convenient form is to relate voltage to *current*, as we have in other parts of this course. Recall that current is the flow of charge per unit time, that is, $i = \frac{dQ}{dt}$. Then,

$$i = \frac{dQ}{dt} = \frac{d}{dt}(Cv) = C\frac{dv}{dt}.$$

The equation

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



is how we normally think of capacitors in electrical engineering, that is, current is proportional to the *change in voltage with respect to time*. Note that v and i are labeled according to the passive sign convention, with current going into the terminal that is labeled positive.

It's worth pausing for a moment to consider what this means, and what it does *not* mean. The above equation tells us nothing about the voltage across the capacitor itself—only the *rate of change* of voltage. If (and only if) positive current is flowing into the positive terminal, then the capacitor's voltage is *increasing*. If (and only if) positive current is flowing into the negative terminal, then the capacitor's voltage is *decreasing*. If (and only if) there is no current through the capacitor, the capacitor's voltage is *constant*.

Capacitor in steady state looks like open circuit

It's worth pointing out two corollaries of the equation $i = C\frac{dv}{dt}$. First, if the circuit has reached a state where everything is steady, then (by assumption) $\frac{dv}{dt} = 0$, which implies that $i = 0$, independent of C. We often phrase this as the maxim, *at steady state, capacitors look like an open circuit.* We'll come back to this in Section 6.3.

No instantaneous voltage change

Another corollary of the equation $i = C\frac{dv}{dt}$ is that *the voltage across a capacitor can't change instantaneously.* A sudden change would require $\frac{dv}{dt} = \infty$, implying that $i = \infty$, which isn't physically possible. This fact—that even if something else suddenly changes, the voltage across a capacitor cannot—will become useful in our study of transient response shortly.

More generally, since rapid changes in voltage require large currents, one good way to think about capacitors is that they tend to "resist" changes in voltage, or that they'll try to "hold" the voltage for a little bit. The larger the capacitance, the more they tend to do so.

Note that, in capacitors, this observation only applies to voltage. There's nothing preventing the current through a capacitor from changing abruptly.

Capacitors are linear

The relationship between i and v for a capacitor is still linear: if you double v (for all time), i doubles with it. It's a bit hard to imagine, given that i and v aren't really directly related, but recall that constant multiplicative coefficients survive differentiation, which is a property of linear systems.[6] Because capacitors are linear, a lot of circuit analysis techniques we studied with resistors still work with capacitors, most notably superposition.

[6] We don't study linearity formally in this course, but you will in EE 102A.

If you think of charge as a fluid, then you can think of a capacitor like a large tank.[7] The height of the water in the tank represents the voltage on the capacitor. While we can instantaneously turn the flow of water off and on (changing the current), we cannot instantaneously change the water level of this tank. Similarly, we cannot instantaneously change the voltage across the capacitor. Its value changes as a result of the integration of the current being added to the tank/capacitor, and the value of the capacitance is related to the area of the tank.

Energy stored in capacitor

Because capacitors are energy storage devices, we might also want to know how much energy is stored in a capacitor. Recall that power is the rate of change of energy with respect to time. Therefore, we can solve for the power P, substitute values from the equations above, and integrate over time to get energy:

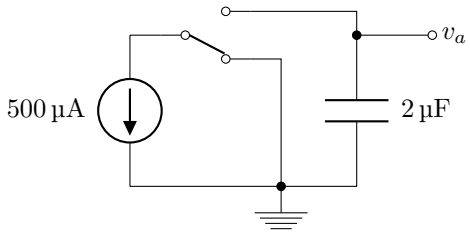
$$E = \int P\,dt = \int iv\,dt = \int C\frac{dv}{dt}v\,dt = \int_0^V Cv\,dv = \frac{1}{2}CV^2,$$

that is,

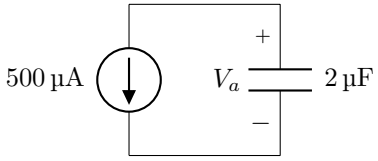
$$E = \frac{1}{2}CV^2, \tag{6.3}$$

where V is the voltage across the capacitor now.

Example 6.1 In the (contrived) circuit below, at $t = 0\,\text{ms}$, $v_a = 10\,\text{V}$. The switch begins on the lower throw, and moves to the upper throw at $t = 0\,\text{ms}$. When does the capacitor's voltage reach 0 V?



Solution 6.1:
By passive sign convention, the current through the capacitor is negative, thus the capacitor is discharging:



[7] You need to be careful with this analogy, since there is plus and minus charge, and there is not negative water. When water flows into a tank to fill it up, it only flows into the top of the tank. With a capacitor there are really two tanks. Charge flows into the top tank, which starts filling it up, but an equal amount of charge also leaves the bottom tank, filling the bottom tank with negative charge. A better fluid analogy would be to model a capacitor as two tanks. The first one is right side up and empty, and the second is upside down and full. When liquid flows into the top tank, to start to fill it up, and equal amount of liquid flows out of the bottom tank, and is replaced by bubbles.

Current is given by

$$i = -C\frac{dV}{dt}$$

The rate of change of voltage is given by

$$\frac{dV}{dt} = -\frac{i}{C} = -\frac{500\,\mu\text{A}}{2\,\mu\text{F}} = -250\,\text{V/s}$$

Thus, the amount of time for the voltage across the capacitor to reach 0 V is

$$\frac{10\,\text{V}}{250\,\text{V/s}} = 40\,\text{ms}$$

6.3 Capacitors in Steady State

If a circuit has reached a state where all currents and voltages have "settled" into a constant, long-term state, we say that the circuit is in *steady state*. That is, in steady state, $\frac{di}{dt} = 0$ and $\frac{dv}{dt} = 0$ for every current and voltage in the circuit. As far as capacitors go, this means that

$$i = C\frac{dv}{dt} = C \cdot 0 = 0,$$

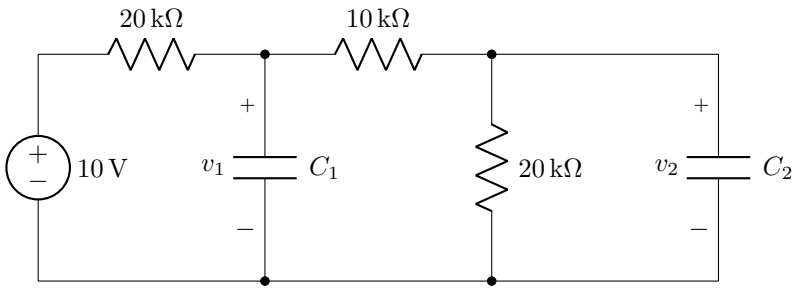
Capacitors in steady state

independent of C. If the current through a device is identically zero, then it looks the same as an open circuit. This gives rise to a maxim on capacitors in steady state:

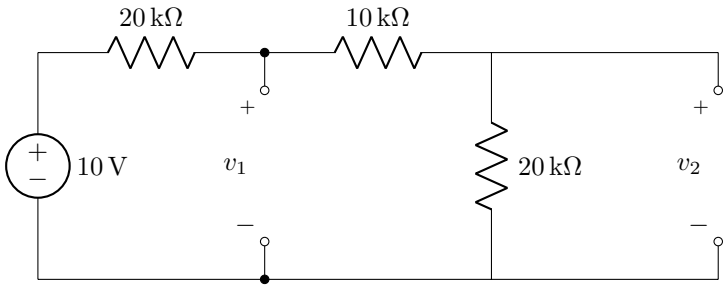
In steady state, capacitors behave like open circuits.

We illustrate this with an example.

Example 6.2 The circuit below has reached steady state. Find v_1 and v_2.



Solution 6.2:
Since the circuit is in steady state, all capacitors look like open circuits. Redraw the circuit accordingly:



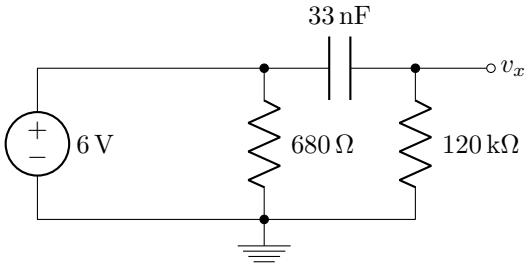
Now this is just a resistor network, so using the voltage divider rule,

$$v_1 = 10\,\text{V}\left(\frac{20\,\text{k}\Omega + 10\,\text{k}\Omega}{20\,\text{k}\Omega + 20\,\text{k}\Omega + 10\,\text{k}\Omega}\right) = 6\,\text{V},$$
$$v_2 = 10\,\text{V}\left(\frac{20\,\text{k}\Omega}{20\,\text{k}\Omega + 20\,\text{k}\Omega + 10\,\text{k}\Omega}\right) = 4\,\text{V}.$$

Example 6.3 The circuit below has reached steady state.

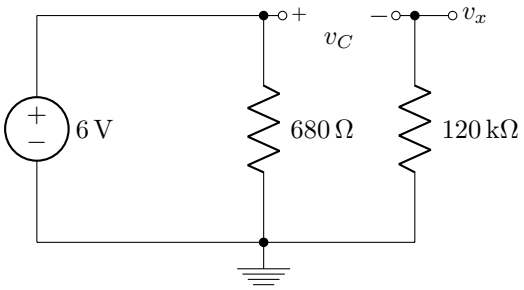
(a) What is the voltage v_x?

(b) What is the voltage across the capacitor? (Be sure to specify the direction.)



Solution 6.3:

In the steady state, capacitors look like open circuits so we can simplify the circuit to



(a) No current flows through the 120 kΩ resistors, so v_x is effectively connected to ground, $v_x = 0\,\text{V}$.

(b) The voltage across the capacitor is $V_C = 6\,\text{V} - v_x = 6\,\text{V} - 0\,\text{V} = 6\,\text{V}$.

6.4 Uses of Capacitors in Circuits

Capacitors, like many electrical components, come in a variety of shapes and sizes. The physical size of a capacitor generally depends on the amount of energy it can store. Not surprisingly capacitors which can store more energy are larger. This means capacitors with larger capacitance, or higher voltage ratings are larger (remember the energy stored is CV^2 so doubling the max voltage increases the energy by 4x). Electrolytic capacitors have the largest capacitors per unit volume, but they have limited voltage and as polarized, as described before. In some circuits we add explicit capacitors to help the circuits function, while in other situations we add capacitors to our schematic, to model physical effects encountered on a real circuit board.

Capacitors to Control Supply Voltage

We often use capacitors in our circuit to help keep the power supply voltage constant. For this we use large capacitors, like we did in our smart useless box. One would think that this would not be necessary, since voltage sources are supposed to supply a constant voltage independant of current. But all real sources are not idea, and have some voltage changes with current. As you draw more current, especially if the change is sudden, voltage will drop. We can mitigate these effects by connecting a capacitor between Vdd and Gnd. We know capacitors resist changes in voltage, so the capacitor will work to keep Vdd constant. Essentially, the capacitor acts as a energy reserve for the circuit, supplying energy when the demands of the circuit exceed the battery's (or other voltage source's) capabilities.

Tone Dependent Circuits

In addition to adding capacitors to help keep the power supply stable, capacitors are most often used to create a circuit where its behavior is different for different input tones. Here we use the fact that the current through a capacitor depends on how fast the voltage is changing to help us. We know that high-frequency tones will change faster than low-frequency tones. This means a circuit with a capacitor in it might treat the two tones differently (and it does). How to use capacitors for this type of circuit is described in Chapter 7.

Capacitors in MOS Transistors

Chapter 4 described a MOS transistor, and explained that the gate terminal controlled whether current flowed between the source and drain, but no current ever flowed into the gate terminal. What it failed to explain is that the gate terminal of the MOS transistor is really one terminal of a capacitor. The other terminal of the capacitor is the source terminal of the MOS transistor. This means a complete model of a MOS transistors should include this "gate" capacitance, in addition to the controlled resistance between the source and drain. Adding this capacitor to our model will lead to delay in our logic gates, as we will show later in this chapter.

This capacitor also explains why you need to use a pull up resistor when you connect a switch

to the input of a MOS gate. From the equation of a capacitor

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

we see that if $i = 0$, then $\frac{dv}{dt} = 0$. As a result, if there is no current, the voltage across the capacitor (from the gate to the source) remains constant. For example, if we simply disconnect the gate terminal from Gnd, without driving it to another voltage, it will remain at Gnd. Thus, we need a current to cause a change in voltage and change the gate voltage of the MOS transistor.

Real Wires

All real wires also have capacitance. As we saw above, this means wires require some charge to change their voltage. Remember that voltage is defined as potential energy per charge, so this observation makes sense in the context of our definition. This means that any time we want to change the voltage of a wire, we need charge to flow into it. The amount of charge we need to produce a given change of voltage is governed by the equation:

$$Q = Cdv.$$

We're starting to see a theme. Capacitance governs how fast we can change voltages and the energy we need to do so. This result governs the speed and power consumption of modern electronics, including your computers!

6.5 Transient Response of RC Circuits

If a circuit has only one capacitor (or capacitors that can be reduced using series/parallel rules to one capacitor), we say it is a *first-order* circuit. In many applications, these circuits respond to a *sudden change* in an input: for example, a switch opening or closing, or a digital input switching from low to high. Just after the change, the capacitor takes some time to charge or discharge, and eventually settles on its new steady state. We call the response of a circuit immediately after a sudden change the *transient response*, in contrast to the steady state.

Recall that in a capacitor, $i = C\frac{dv}{dt}$. If a capacitor experienced a sudden change in voltage—even a small sudden change—that would be tantamount to $\frac{dv}{dt} = \infty$, which in turn requires $i = \infty$. This isn't physically possible—you can have large currents, but not infinite ones. This leads to another maxim about capacitors:

No instantaneous voltage change

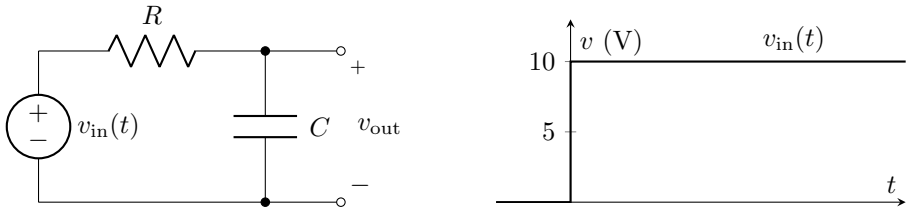
The voltage across a capacitor can't change instantaneously.

Note that this constraint only relates to the *voltage* across a *capacitor*. The current through a capacitor is free to change as suddenly as it likes, and voltages across other devices can in principle change instantaneously. But the voltage across a capacitor is an important means of relating voltage and current in a circuit just *before* a sudden change, to that just *after* a sudden change.

To distinguish between "just before" and "just after", we often use $-$ and $+$ superscripts next to the time in question. So, for example, $v(0^-)$ means the voltage $v(t)$ just before $t = 0$, and $v(0^+)$ means the voltage $v(t)$ just after $t = 0$.

6.5.1 A First Example

To get us started, consider the circuit in Figure 6.5, whose behavior we'll derive from first principles. The voltage source provides $v_{\text{in}}(t) = 0$ for $t < 0$, and $v_{\text{in}}(t) = 10\,\text{V}$ for $t \geq 0$. We wish to find $v_{\text{out}}(t)$.



(a) circuit diagram (b) graph showing $v_{\text{in}}(t)$

Figure 6.5: Basic RC charging circuit

First, a few observations, using the steady state analysis we discussed in Section 6.3.

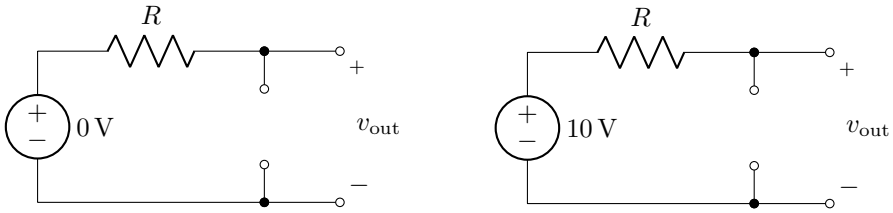
- *Just before* the step in v_{in} from 0 V to 10 V at $t = 0$, the input has (by assumption) been constant for a very long time, so the circuit should have reached steady state. Therefore, the capacitor looks like an open circuit, as shown in Figure 6.6(a). Since there's no current through the resistor, $v_{\text{out}}(0^-) = v_{\text{in}}(0^-) = 0\,\text{V}$.

- *Just after* the step, $v_{\text{out}}(0^+)$ must be the same, $v_{\text{out}}(0^+) = v_{\text{out}}(0^-) = 0\,\text{V}$, because it so happens in this case that v_{out} is the voltage across a capacitor, which can't change instantaneously.

- *Long after* the step, if we wait long enough the circuit will reach steady state, then $v_{\text{out}}(\infty) = 10\,\text{V}$, as shown in Figure 6.6(b).

What happens in between? Using Kirchoff's current law applied at the top-right node, we could write

$$\frac{10\,\text{V} - v_{\text{out}}}{R} = C\frac{dv_{\text{out}}}{dt}.$$



(a) $t = 0^-$ (just before $t = 0$) (b) $t = \infty$ (final steady state)

Figure 6.6: Circuit of Figure 6.5 in steady state

Solving this differential equation yields

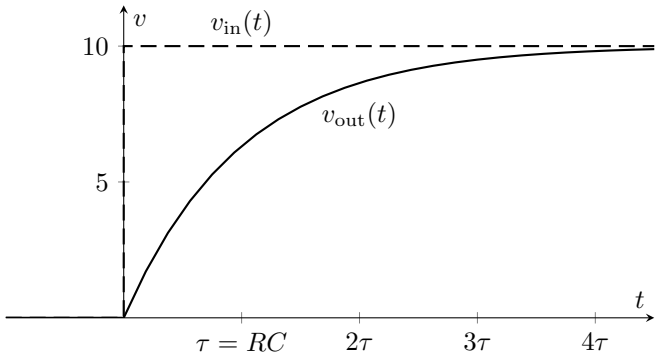
$$\begin{aligned}
\frac{1}{RC}dt &= \frac{1}{10\,\mathrm{V} - v_{\mathrm{out}}}dv_{\mathrm{out}} \\
\frac{1}{RC}\int dt &= \int \frac{1}{10\,\mathrm{V} - v_{\mathrm{out}}}dv_{\mathrm{out}} \\
\frac{t}{RC} &= -\ln(10\,\mathrm{V} - v_{\mathrm{out}}) + c \\
e^{\frac{-t}{RC}+c} &= 10\,\mathrm{V} - v_{\mathrm{out}} \\
10\,\mathrm{V} - Ae^{\frac{-t}{RC}} &= v_{\mathrm{out}}, \qquad \text{(where } A = e^c\text{)}
\end{aligned}$$

and applying the initial condition $v(0^+) = 0$ that we found above,

$$Ae^{\frac{-0}{RC}} = 10\,\mathrm{V} - 0 \quad \Longrightarrow \quad A = 10\,\mathrm{V}$$

So the solution to the differential equation is

$$v_{\mathrm{out}}(t) = 10 - 10e^{-\frac{t}{RC}}.$$



What's happening? Immediately after the step, the current flowing through the resistor—and hence the capacitor (by KCL)—is $i(0^+) = \frac{10\,\mathrm{V}}{R}$. Since $i = C\frac{dv_{\mathrm{out}}}{dt}$, this current causes v_{out} to start rising. This, in turn, *reduces* the current through the resistor (and capacitor), $\frac{10\,\mathrm{V}-v_{\mathrm{out}}}{R}$. Thus, the rate of change of v_{out} decreases as v_{out} increases. The voltage $v_{\mathrm{out}}(t)$ technically never reaches steady state, but after about $3RC$, it's very close.

6.5.2 Transient Response Equation

It turns out that *all* first-order circuits respond to a sudden change in input with some sort of exponential decay, similar to the above. Therefore, we don't solve differential equations every time we see a capacitor, and we won't ask you to solve any.

Instead, we use the following shortcut: In any first-order circuit, if there is a sudden change at $t = 0$, the transient response for a voltage is given by

Transient response equation

$$v(t) = v(\infty) + [v(0^+) - v(\infty)]e^{-t/\tau}, \tag{6.4}$$

where $v(\infty)$ is the (new) steady-state voltage; $v(0^+)$ is the voltage just *after* time $t = 0$; τ is the *time constant*, given by $\tau = RC$ (more about this in Section 6.5.2, and in both cases R is the *resistance seen by the capacitor* (explained below).

The transient response for a current is the same, with $i(\cdot)$ instead of $v(\cdot)$:

$$i(t) = i(\infty) + [i(0^+) - i(\infty)]e^{-t/\tau}. \tag{6.5}$$

The voltage or current in question needn't be that of a capacitor. So long as the circuit is a first-order circuit, any voltage or current will follow this template.

How to find R in $\tau = RC$

What do we mean by the "resistance seen by the capacitor"? Informally, it means the resistance you would think the rest of the circuit had, if you were the capacitor. More precisely, you find it using these steps:

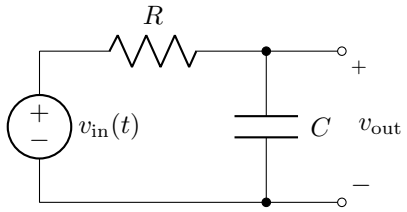
1. Zero out all sources (*i.e.* short all voltage sources, open all current sources)
2. Remove the capacitor
3. Find the resistance of the resistor network whose terminals are where the capacitor was

Workflow for transient response

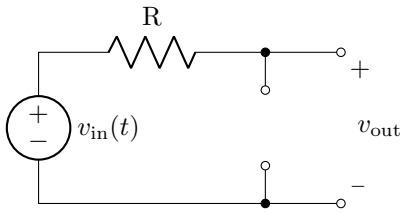
Typically, your workflow for finding an expression for $v(t)$ in the transient response of a circuit would look something like this (replacing v with i if you're finding a current):

1. Establish what $v(0^-)$ is.
2. Use $v(0^-)$ to establish what $v(0^+)$.
3. Use steady-state analysis to establish what $v(\infty)$ is.
4. Find the resistance seen by the capacitor (using the steps above).
5. Find the time constant, $\tau = RC$.
6. Substitute $v(0^+)$, $v(\infty)$ and τ into the transient response equation (6.4) or (6.5).

Example 6.4 Consider the following circuit, whose voltage source provides $v_{\text{in}}(t) = 0$ for $t < 0$, and $v_{\text{in}}(t) = 10\,\text{V}$ for $t \geq 0$. Use the first-order transient response equation to find the result for $v_{\text{out}}(t)$



Solution 6.4:
Let's first find the steady state voltage, $v_{\text{out}}(\infty)$. In steady state, capacitors look like open circuits, thus we can simplify the circuit to



No current is flowing, thus there is no voltage drop across the resistor and $v_{\text{out}} = v_{\text{in}} = 10\,\text{V}$. Since $v_{\text{out}}(t)$ is the voltage across a capacitor, it can't change instantaneously, so the voltage just after time $t = 0$ is equal to the voltage just before $t = 0$, $v(0^+) = v(0^-) = 0$. We plug these values for $v_{\text{out}}(0^+)$ and $v_{\text{out}}(\infty)$ into the equation

$$v_{\text{out}}(t) = v_{\text{out}}(\infty) + \left[v_{\text{out}}(0^+) - v_{\text{out}}(\infty)\right] e^{\frac{-t}{\tau}} = 10\,\text{V} + [0\,\text{V} - 10\,\text{V}]\, e^{\frac{-t}{\tau}}$$
$$v_{\text{out}}(t) = 10\,\text{V} - 10\,\text{V} \cdot e^{\frac{-t}{\tau}} = 10\,\text{V} - 10\,\text{V} \cdot e^{\frac{-t}{RC}}$$

Charging a capacitor

In the special case where the capacitor starts fully discharged (no voltage across it) and charges up to a final voltage V, we have $v(0^+) = 0$ and $v(\infty) = V$. Then the transient response equation (6.4) reduces to

$$v(t) = V - Ve^{-t/\tau} = V(1 - e^{-t/\tau}). \tag{6.6}$$

We sometimes call (6.6) the *charging equation*, reflecting that it starts at zero and charges towards the final steady-state value. The example in Section 6.5.1 is an example of this case, where $V = 10\,\text{V}$.

Discharging a capacitor

In the special case where a capacitor starts at an initial voltage V and discharges down to zero, we have $v(0^+) = V$ and $v(\infty) = 0$. Here, (6.4) reduces to

$$v(t) = Ve^{-t/\tau} \tag{6.7}$$

Unsurprisingly, we sometimes call (6.7) the *discharging equation*.

The parameter τ (the Greek letter *tau*) is known as the *time constant* of the circuit. It has units of seconds (you should verify this for yourself), and it governs the "speed" of the transient response. Circuits with higher τ take longer to get close to the new steady state. Circuits with short τ settle on their new steady state very quickly.

More precisely, every time constant τ, the circuit gets $1 - e^{-1} \approx 63\%$ of its way closer to its new steady state. This is illustrated in Figure 6.7. Memorizing this fact can help you draw graphs involving exponential decays quickly.

After 3τ, the circuit will have gotten $1 - e^{-3} \approx 95\%$ of the way, and after 5τ, more than 99%. So, after a few time constants, for practical purposes, the circuit has reached steady state. Thus, the time constant is itself a good rough guide to "how long" the transient response will take.

Of course, mathematically, the steady state is actually an asymptote: it never *truly* reaches steady state. But, unlike mathematicians, engineers don't sweat over such inconsequential details.

6.5.3 CMOS Gate Delay

We can use our understanding of the delay in circuits with resistors and capacitors to find the delay of CMOS logic gates. Let's examine the inverter shown in Figure 6.8. These transistors drive a wire, which we know has a small amount of capacitance.

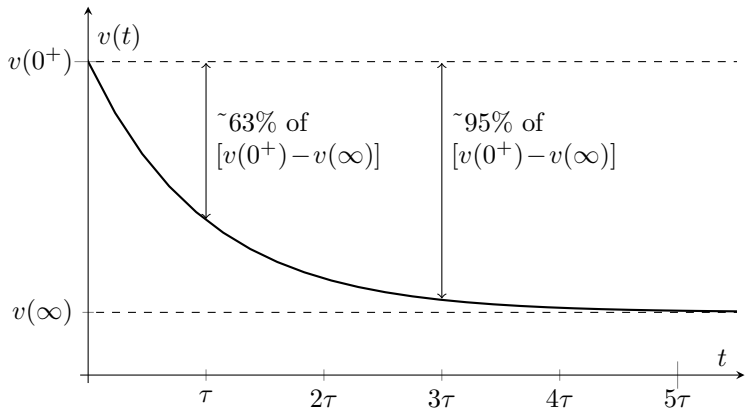


Figure 6.7: Time constant

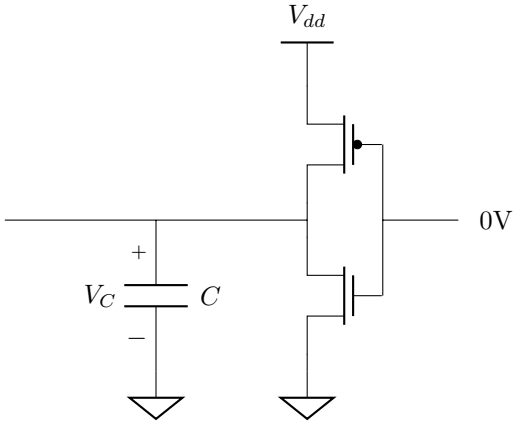
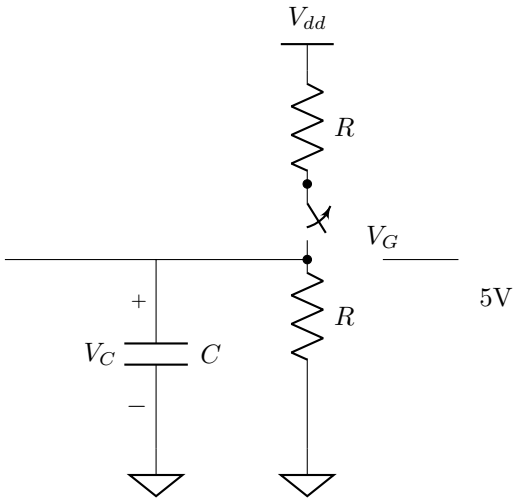


Figure 6.8: Simple schematic of an inverter showing a capacitor on the output. This capacitor is modeling the wire capacitance, and the gate capacitance of any MOS transistors that this output drives.

We can model the transistors as voltage dependent switches in series with a resistor. Since we are driving the input to 0V, the pMOS transistor will be on and the nMOS transistor will be off. Assuming this circuit has been in this state for a sufficiently long time, we know that $V_C = V_{dd}$. At $t = 0^+$, let drive the gate voltage V_G to V_{dd} so that the pMOS transistor turns off and the nMOS transistor turns on:



This is just the capacitor discharge problem that we previously solved. We know that the output voltage will fall from Vdd to Gnd, as a decaying exponential, and the time constant of that exponential will be RC:

$$V_C(t) = V_{dd} e^{\frac{-t}{RC}}$$

It is this RC delay that sets the speed of CMOS gates. To achieve the very high performance we have today (gate delay around 10-20 ps) requires very small load capacitance on each gate, and relatively low transistor resistance.

6.5.4 Examples with Several Capacitors, Resistors or Sources

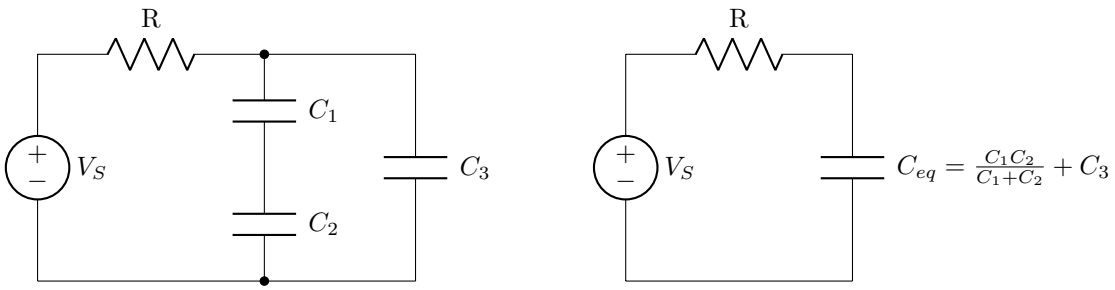
RC circuits with several capacitors

If there are **multiple capacitors** in the circuit, you might be able to reduce them to a single capacitor using series and parallel rules. If you can, then it's still a first-order circuit. Take, for example, the circuit in Figure 6.9(a). Using the series and parallel rules from Equations (6.8) and (6.9), we can reduce the three capacitors C_1, C_2 and C_3 to a single equivalent capacitor of capacitance $C_{eq} = \frac{C_1 C_2}{C_1 + C_2} + C_3$, as shown in Figure 6.9(b).

A word of caution: If you *can't* reduce the capacitors to a single equivalent capacitor (say, because they're not actually in series or parallel), then you probably don't have a first-order circuit, and you can't apply the transient response equation (which only applies to first-order circuits).

RC circuits with several resistors or sources

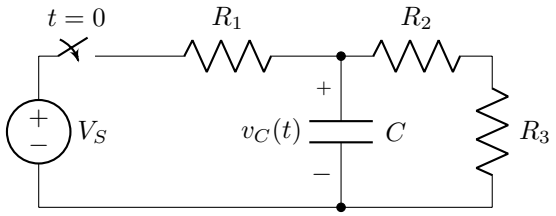
If there are **multiple resistors or sources** in the circuit, the workflow to apply the transient response equation (6.4) remains the same. Simply find $v(0^+)$, $v(\infty)$ and $\tau = RC$ using the same methods. We illustrate with two examples.



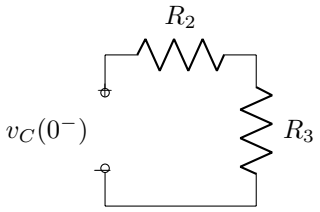
(a) original circuit (b) equivalent circuit

Figure 6.9: RC circuit with several capacitors

Example 6.5 The switch in the circuit below, having been open for a long time, closes at $t = 0$. Find an expression for $v_C(t)$, the voltage across the capacitor in this circuit.

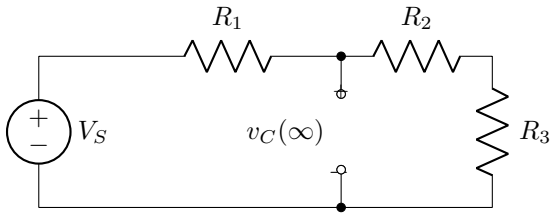


Solution 6.5:
Before the switch closes, the switch having been open for a long time, the circuit is in steady state. So the circuit reduces to the following:



There's no current through the resistors, so $v_C(0^-) = 0$. Since voltage through a capacitor can't change instantaneously, we have $v_C(0^+) = 0$.

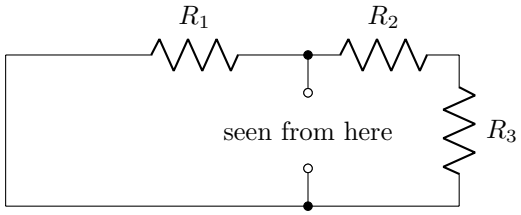
After the switch closes, the eventual steady state of the circuit is as follows:



This is just a voltage divider, so

$$v_C(\infty) = V_S \left(\frac{R_2 + R_3}{R_1 + R_2 + R_3} \right).$$

Finally, to find the resistance seen by the capacitor during $t > 0$, note that the switch is closed for $t > 0$, and short the voltage source:



Seen from where the capacitor was, we see R_1 in parallel with the R_2–R_3 series combination, yielding

$$R_{eq} = R_1 \parallel (R_2 + R_3) = \frac{(R_2 + R_3)R_1}{(R_2 + R_3) + R_1},$$

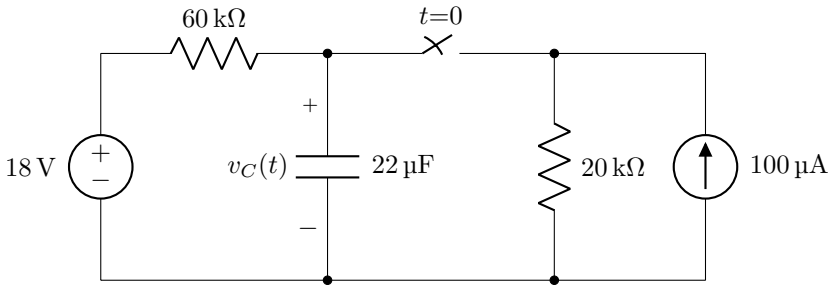
so the time constant is

$$\tau = R_{eq}C = \frac{(R_2 + R_3)R_1 C}{(R_2 + R_3) + R_1}.$$

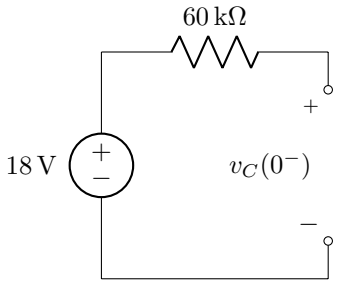
Putting it all into the transient response equation, we have

$$\begin{aligned} v(t) &= v(\infty) + [v(0^+) - v(\infty)]e^{-t/\tau} \\ &= V_S \left(\frac{R_2 + R_3}{R_1 + R_2 + R_3} \right) + \left[0 - V_S \left(\frac{R_2 + R_3}{R_1 + R_2 + R_3} \right) \right] e^{-t/R_{eq}C} \\ &= V_S \left(\frac{R_2 + R_3}{R_1 + R_2 + R_3} \right) \left[1 - \exp \left(-t \frac{(R_2 + R_3) + R_1}{(R_2 + R_3)R_1 C} \right) \right]. \end{aligned}$$

Example 6.6 The switch in the circuit below has been open for a long time and closes at $t = 0$. Find $v_C(t)$ for $t > 0$.

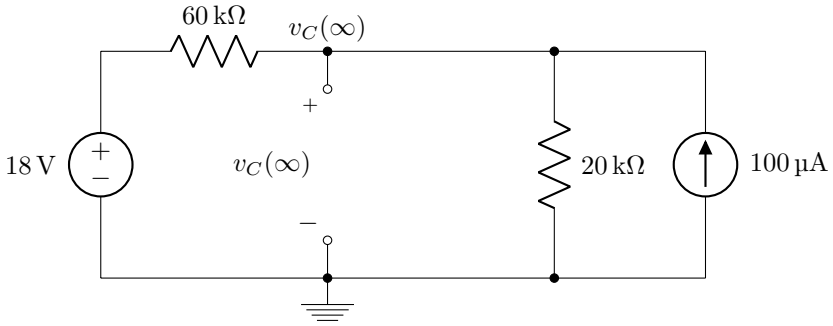


Solution 6.6:
To find $v_C(0^+)$, note that $v_C(0^-)$ is the voltage across a capacitor, so $v_C(0^+) = v_C(0^-)$. Then, what is $v_C(0^-)$? Just before $t = 0$, the switch has been open for a long time, so the circuit would have been in steady state with the switch open:



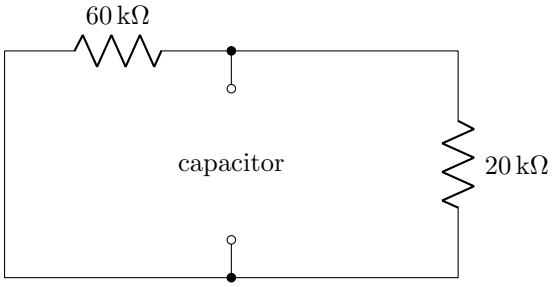
No current can flow through that resistor (there's an open circuit in the way), so the voltage across the resistor is zero, then $v_C(0^+) = v_C(0^-) = 18\,\text{V}$.

Next, to find $v_C(\infty)$, find the steady-state v_C when the switch is closed. Label the node at the bottom ground, and then use nodal analysis on the node on the other side of the capacitor:



$$\frac{v_C(\infty) - 18\,\text{V}}{60\,\text{k}\Omega} + \frac{v_C(\infty)}{20\,\text{k}\Omega} = 100\,\mu\text{A}$$
$$v_C(\infty) = 6\,\text{V}.$$

To find the resistance seen by the capacitor, short the voltage source and open the current source:



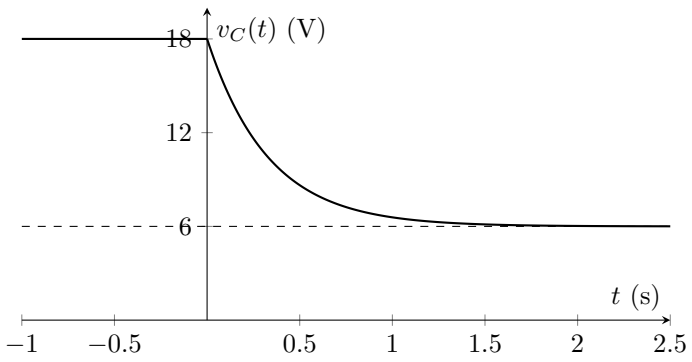
The resistance seen by the capacitor is

$$60\,\text{k}\Omega \parallel 20\,\text{k}\Omega = \frac{60\,\text{k}\Omega \cdot 20\,\text{k}\Omega}{60\,\text{k}\Omega + 20\,\text{k}\Omega} = 15\,\text{k}\Omega.$$

The time constant is then $\tau = 15\,\text{k}\Omega \times 22\,\mu\text{F} = 330\,\text{ms}$. Putting this all together,

$$\begin{aligned} v(t) &= v(\infty) + [v(0^+) - v(\infty)]e^{-t/\tau} \\ &= 6 + (18-6)e^{-t/330\,\text{ms}} \quad (\text{V}) \\ &= 6 + 12e^{-t/330\,\text{ms}} \quad (\text{V}), \qquad t > 0. \end{aligned}$$

Here's a graph of $v_C(t)$:



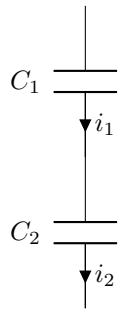
6.6 Capacitors in Series and Parallel

We would like to create equations to simplify multiple capacitors in a circuit. To do this it would be nice to see if there was an equivalent "resistance" for a capacitor. We can't really find one, since the current depends on ΔV not V, but we can say that $\Delta V = i\frac{\Delta T}{C}$. This is interesting in two ways.

First is says that the effective resistance is related to $1/C$, so circuits where R adds will need to combine $1/C$. Second it shows how the resistance depends on the ΔT. If ΔT is small, the effective resistance will be small, but if the time is large, the resistance will also be large.

6.6.1 Capacitors in Series

Like with resistors, when we have multiple capacitors in a circuit, we often want to replace them with one equivalent capacitor. We'll start by analyzing capacitors in series:



Using our equation for the current through a capacitor, we have $i_1 = C_1 \frac{dV_1}{dt}$ and $i_2 = C_2 \frac{dV_2}{dt}$ where V_1 and V_2 are the voltages across C_1 and C_2 respectively. We want to find one capacitor with capacitance C_{eq} such that the voltage across it is $V_1 + V_2 = V_{tot}$ and the current through it is $i = i_1 = i_2$. We know from KCL that $i_1 = i_2$. With this information, we now rearrange the equations to get:

$$i_1 = C_1 \frac{dV_1}{dt} \qquad\qquad i_2 = C_2 \frac{dV_2}{dt}$$
$$\frac{i_1}{C_1} = \frac{dV_1}{dt} \qquad\qquad \frac{i_2}{C_2} = \frac{dV_2}{dt}$$

$$i\left(\frac{1}{C_1} + \frac{1}{C_2}\right) = \frac{dV_{tot}}{dt}$$
$$i = \left(\frac{1}{\frac{1}{C_1} + \frac{1}{C_2}}\right) \frac{dV_{tot}}{dt}$$

Capacitors in series

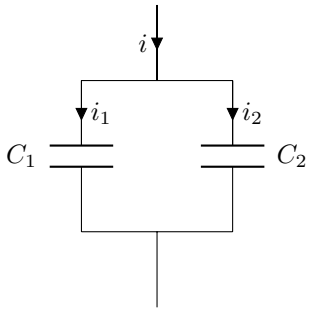
This looks a lot like our equation for current through a capacitor. Thus, we have the following relation for capacitors in series:

$$\frac{1}{C_{eq}} = \frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n}. \tag{6.8}$$

This makes sense intuitively, as both capacitors will be experiencing the same current, and the voltage across both will increase with respect to this current. Thus, the total voltage across both capacitors will increase at a greater rate than either of the voltages accross individual capacitors. We also notice that this relation looks like the relation for resistors in parallel.

6.6.2 Capacitors in Parallel

We also want to be able to replace capacitors in parallel. Let's analyze the following:



By KVL, we know that the voltage across C_1 must equal the voltage across C_2. Call this voltage V. Also, by KCL, we know that $i = i_1 + i_2$. We substitute the currents through each capacitor:

$$\begin{aligned} i &= i_1 + i_2 \\ &= C_1 \frac{dV}{dt} + C_2 \frac{dV}{dt} \\ &= (C_1 + C_2) \frac{dV}{dt} \end{aligned}$$

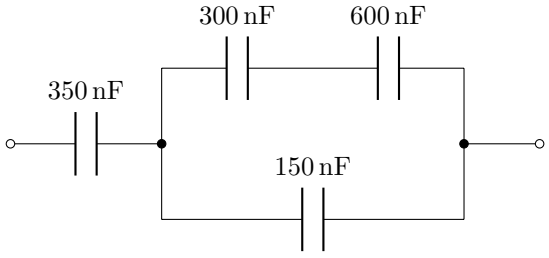
Thus, we have the following relation for capacitors in parallel:

Capacitors in parallel

$$C_{eq} = C_1 + C_2 + ... + C_n \tag{6.9}$$

Going back to our water tank analogy, the summation makes sense. Putting two capacitors in parallel is like putting two water tanks in parallel. We also notice that this relation is similar to resistors in series. In this section, we have seen that the equivalent capacitance equations are the same as the equivalent resistance ones, but the series and parallel behaviors are flipped.

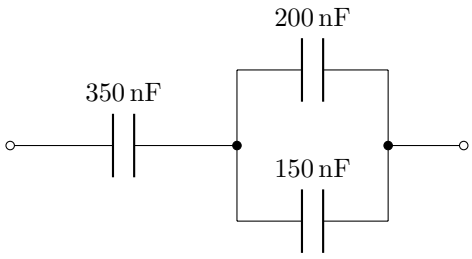
Example 6.7 You wish to replace the following network of capacitors with a single capacitor. What should its value be?



Solution 6.7:
We first find the equivalent capacitance of the 300 nF and 600 nF in series

$$C_s = \frac{C_1 C_2}{C_1 + C_2} = \frac{300\,\text{nF} \times 600\,\text{nF}}{300\,\text{nF} + 600\,\text{nF}} = 200\,\text{nF}$$

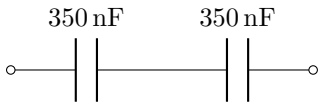
The circuit can now be simplified to



We now find the equivalent capacitance of the 200 nF and 150 nF in parallel,

$$C_p = C_1 + C_2 = 200\,\text{nF} + 150\,\text{nF} = 350\,\text{nF}$$

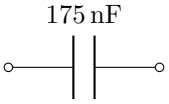
The circuit can now be simplified to



The equivalent capacitance of these two 350 nF capacitors in parallel is given by

$$C_s = \frac{C_1 C_2}{C_1 + C_2} = \frac{350\,\text{nF} \times 350\,\text{nF}}{350\,\text{nF} + 350\,\text{nF}} = 175\,\text{nF}$$

We replace the initial circuit with the a single 175 nF



6.7 Summary

- Capacitors are devices that store energy in electric fields. They can be polarized or non-polarized.
- Voltage across a capacitor and current through it are related by $i = C\frac{dV}{dt}$. (Section 6.2)
- The energy stored in a capacitor is $\frac{1}{2}CV^2$. (Section 6.2)
- Capacitors in series can be replaced by an equivalent capacitor $\frac{1}{C_{eq}} = \frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n}$. (Section 6.6.1)
- Capacitors in parallel can be replaced by an equivalent capacitor $C_{eq} = C_1 + C_2 + ... + C_n$. (Section 6.6.2)
- In circuits that have reached steady-state at DC, meaning that all voltages and currents have settled at a stable value, capacitors have like open circuits. (Section 6.3)
- The voltage across a capacitor can't change instantaneously. (Section 6.5)
- The time constant of a capacitor is $\tau = RC$. Its units are seconds. (Section 6.5)
- The general equation for voltages in circuits with one capacitor (or one equivalent capacitor) is $v(t) = v(\infty) + [v(0^+) - v(\infty)]e^{-t/\tau}$, where $v(0^+)$ is the voltage in question just *after* $t = 0$, where $v(\infty)$ is the steady-state voltage after the change at $t = 0$ happened, and τ is the time constant. (Section 6.5.2)
- If a capacitor is charging to a known voltage from zero, this reduces to $v(t) = V(1 - e^{-t/\tau})$. (Section 6.5.2)
- If a capacitor is discharging from a known voltage to zero, this reduces to $v(t) = Ve^{-t/\tau}$. (Section 6.5.2)
- When you have multiple capacitors in a circuit, you can often combine them into a capacitor with some C_{eq}. (Section 6.5.4)

Chapter 7

Bode Plots, Impedance and Filters

Our goal in this chapter is to find a way to predict the behavior of an RC (resistor-capacitor) or RL (resistor-inductor) circuit in response to any input signal, not just to step inputs. This is a difficult problem, because inductors and capacitors cause integral and derivatives in the circuit equations, and things get really messy, really fast. To avoid this complexity we are going to take advantage of the fact that the derivative (integral) of a sinusoidal waveform is also a sinusoidal waveform. This means if we drive a RC (or RL, or RLC) circuit with a single frequency sine wave input, all node voltages will be a sine wave at the same frequency. Section 7.3 will show how this allows one to create an effective resistance, called *impedance*, to solve this single "tone" problem.

While it is great to have a method to determine the output if the input is a single sine wave, we actually want the output of a signal that is an arbitrary waveform. Fortunately we still can use the same method, by taking advantage of the fact that:

1. Any waveform can be represented by a sum of different frequency sine waves.

2. Since RC circuits are linear, the output of a circuit driven by a sum of sine waves is the sum of the outputs created by looking at each sine wave one at a time.

Thus if we know how the circuit responds to each sine wave, we can create the actual output by adding all these responses together. This output will be given in a different form than the time waveforms of the previous chapter. The output will be presented as the gain (Vout/Vin) of the circuit for each tone (sine wave frequency), rather than a waveform of voltage vs. time. That is the output of this analysis will create a equation or graph that give the amplitude (and sometimes phase) of the output sine wave as a function of the input amplitude and its frequency.

Since the results of this analysis are often displayed as a gain (Vout/Vin) verses frequency in log-log plots, often using a measure called dB, the first two sections of this chapter reviews log-log plots, and introduce the gain measure called dB.

When you finish this chapter, you should be able to:

- Understand what a Bode plot is and how to use it

- Understand how to describe signals in decibels (dB), where is it commonly used and why we use it.

- Describe the relationship between voltage and current using impedance

7.1 Gain (and dB)

In this and later sections, we're going to talk a lot about the "gain" of a circuit, and how that gain changes with frequency. Gain is simply the ratio of output voltage (the voltage you measure in the circuit), to the input voltage driving the circuit. Gain can be defined for any circuit with a designated port for an input signal, and a port to measure the output signal. In this chapter all the circuits we will build only contain passive components (R, C, L) which can't add energy. Thus all the circuits in this chapter will have a gain less than or equal to one. In the next chapter, we'll be constructing circuits with active elements in them which will provide gains larger than one. A gain larger than 1 means that the signal was amplified (i.e., it came out larger); a gain less than one means it was attenuated (came out smaller).

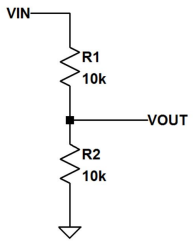
In the real world the range of possible gain is quite large (spanning many orders of magnitude). For this large gain range, measuring the gain on a logarithmic scale can be helpful. Engineers created such a scale, called a "bel"[1]. One "bel" represents a 10X increase in *power*, 2 "bel" represents a 100x increase, and -1 "bel" represents a 10x decrease in power. For reasons unknown to this author, EE's prefer to work in tenths of a bel, which are "decibels" and abbreviated as dB:

$$\text{gain in dB} = 10 \cdot \log_{10} \frac{\text{power out}}{\text{power in}}$$

More often, we're measuring voltage rather than power. Since $P = I \cdot V$ and for a resistor, $I = \frac{V}{R}$, we can also express the dB gain in terms of voltage:

$$P_{in} = \frac{V_{in}^2}{R} \qquad P_{out} = \frac{V_{out}^2}{R} \qquad \text{gain in dB} = 10 \cdot \log_{10} \frac{V_{out}^2}{V_{in}^2} = 20 \cdot \log_{10} \frac{V_{out}}{V_{in}}$$

Finding the gain of a voltage divider



$$Gain_{db} = \frac{V_{out}}{V_{in}} = 20 \cdot \log\left(\frac{10k}{10k+10k}\right) = 20 \cdot \log 0.5 = -6dB$$

Interesting observation: notice that the gain in dB is negative. This is because the gain is 0.5, which is less than 1.

Question: Can the gain in dB of a resistive divider ever be positive? [a]

[a] The gain in dB of a resistive divider can never be positive because the gain is always less than 1, ie. the output is always some fraction of the input.

[1] Yes, it was named after Alexander Graham Bell, and was initially used to measure the signal attenuation caused by the resistance in phone wires.

7.2 Bode plots

Since the gain of circuits with capacitors will depend on frequency, we need a good way to represent this frequency dependence. A *Bode* plot is a very useful method for expressing this information. It simply plots the gain versus the frequency. An example of a Bode plot is shown in Figure 7.1. This plot provides useful information about the function of this circuit. Lower frequencies pass from input to output with out attenuation. But signal tones above 100Hz start being attenuated, and that attenuation grows with frequency. This type of circuit is called a low-pass filter because it passes low frequency tones from the input to the output, but blocks (attenuates) higher frequency tones in the input waveform.

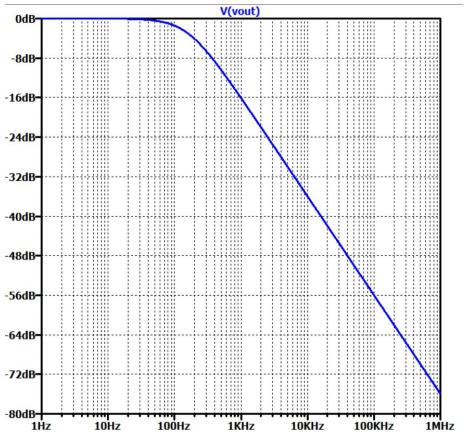


Figure 7.1: Bode plot of a low-pass filter

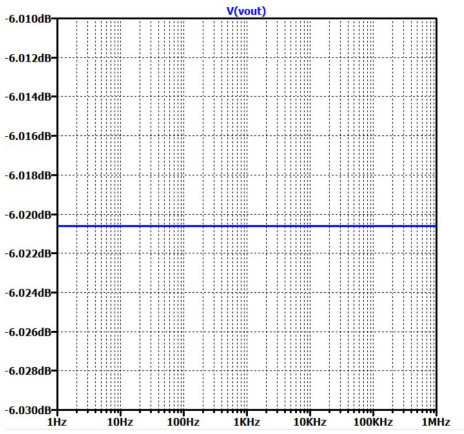


Figure 7.2: Bode plot of the resistive divider

What makes a Bode plot a Bode plot is that both the gain and freq axis are plotted using logarithmic scales.The Y-axis is the gain measured in dB. And while this is plotted on a linear

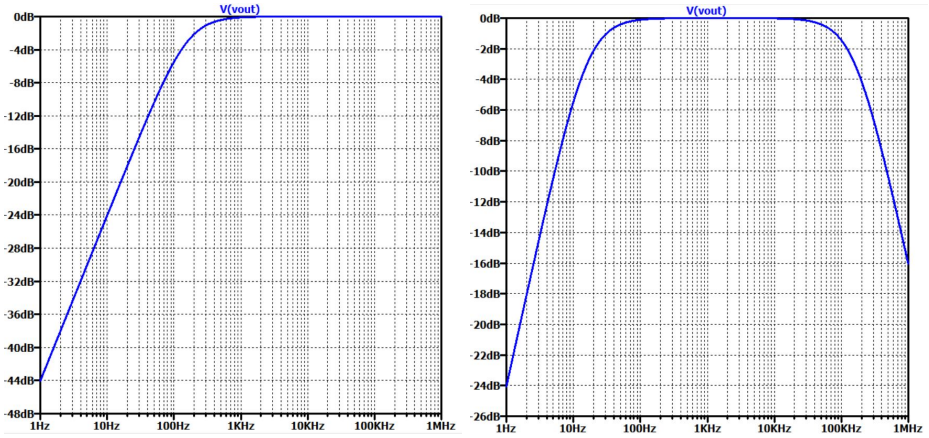
scale, dB is a logarithmic measure, so gain is plotted on a log scale. The X-axis, frequency, is explicitly plotted on a logarithmic scale which results in a log-log plot.

One reason for plotting in this way is because human hearing actually works on a logarithmic scale, which is why you'll often find the unit dB on the specifications for your audio devices. Another is simply that is makes for nice, clean plots from which useful information can be easily extracted, as you will see in Section 7.4.

A Bode plot of a resistive network is relatively boring, since resistance is constant over frequency. Figure 7.2 shows the Bode plot of the resistive divider shown in the prior example box. As you can see, it is simply a flat line at our calculated value of -6dB.

However, when capacitors and inductors are introduced to the circuit, these Bode plots become very useful. By using combinations of R, L, and C, we can create filters, and the Bode plots become much more interesting.

Filters remove unwanted frequencies from electrical signals. Figure 7.3 show the Bode plots of two other types of filters - high-pass filters, which attenuate the low freq tones, and band-pass filters which block both the low frequency and high frequency tones.



(a) Bode plot of a high pass filter (b) Bode plot of a bandpass filter

Figure 7.3: Bode plots of filters

Once we understand how to characterize the impedance of capacitors and inductors, which is discussed in the next section, we will use Bode plots to visualize the effect these devices have on the signals that pass through circuits containing them.

7.3 Generalized Resistance (Impedance)

In previous chapters, we were able to use resistance and Ohm's Law, V= I R, to solve for the voltages and currents in many circuits. Now that we have added capacitors and inductors, we can no longer use Ohm's Law for these components. Wouldn't it be nice, if we could find some effective resistance for these new devices so we can use what we already know to solve circuits that have capacitor and inductors in them? Fortunately this is possible, and this section will show you how to do it. This generalization of resistance is called *impedance*, and is represented by '**Z**'.

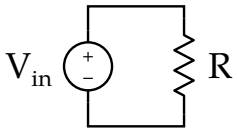
Resistance was defined as the ratio between voltage and current. Since the current through a capacitors and inductors depends on the rate of change of the signal, we can't define this ratio for an arbitrary input, but we can define it when the voltage across the device is sinusoidal. Thus we're going to observe the current when we put a sinusoidal voltage signal across the device. The voltage across the device is a function of time, given by the equation

$$V_{in} = \sin(2\pi f \cdot t)$$

Here, f is the frequency of the signal in Hz, and t is time.

7.3.1 Resistors

Let's start with a really simple circuit, with just the voltage source and a resistor:

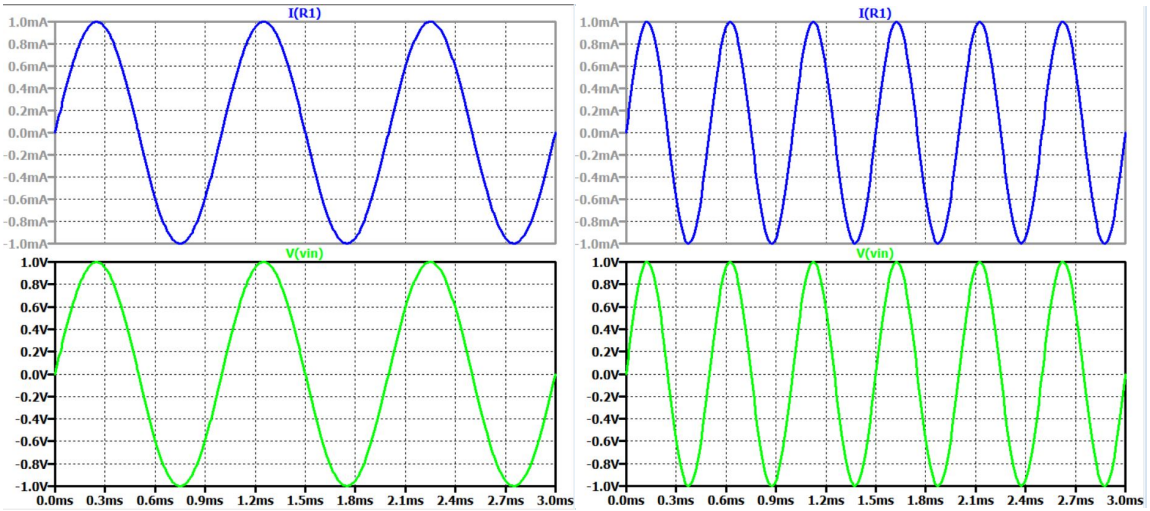


The current through the resistor is given by Ohm's law:

$$i = \frac{V}{R}$$

$$i = \frac{\sin(2\pi f \cdot t)}{R}$$

In the following waveforms we plot the voltage and current across the resistor for R = 1kΩ, and frequencies of 1kHz and 2kHz. Notice that the amplitude of the current is constant over varying frequency.



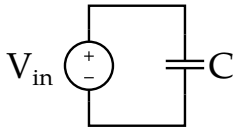
(a) Frequency = 1kHz

(b) Frequency = 2kHz

Figure 7.4: V and I on resistor

7.3.2 Capacitors

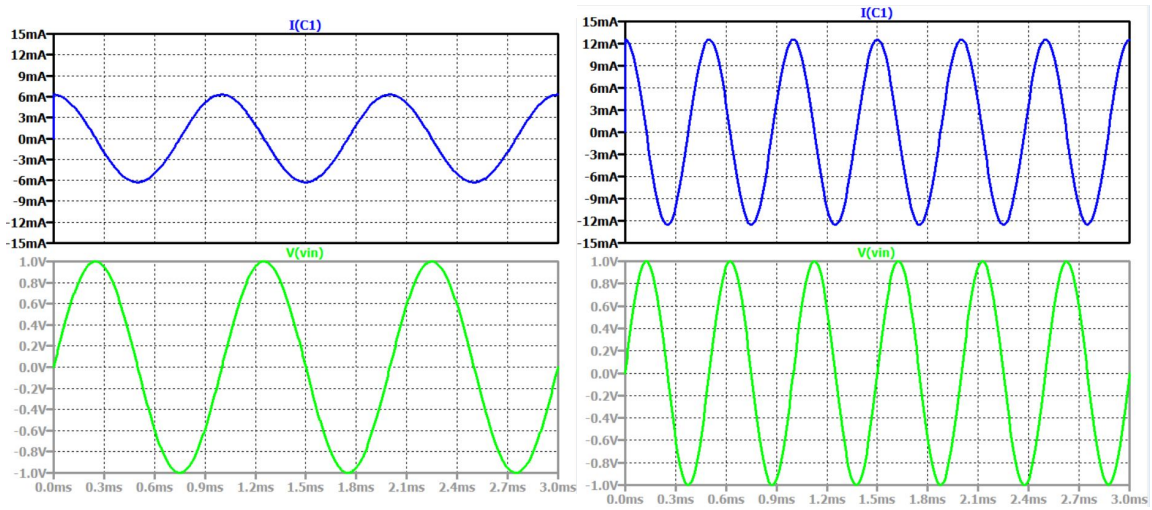
Now let's try a capacitor:



The current is given by the capacitor equation:

$$i = C\frac{dV_{in}}{dt}$$

$$i = C \cdot 2\pi f \cos(2\pi f \cdot t)$$



(a) Frequency = 1kHz, I_{max} = 6.3mA

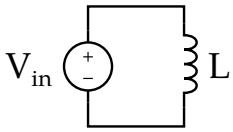
(b) Frequency = 2kHz, I_{max} = 12mA

Figure 7.5: V and I on a 1 µF capacitor

Notice that while the current always has exactly the same frequency as the voltage signal, the amplitude can be different. At low frequencies, the capacitor has very little current flowing through it, as if it were a large resistor. At high frequencies, larger amounts of current flow, as if the resistance is now smaller.

7.3.3 Inductors

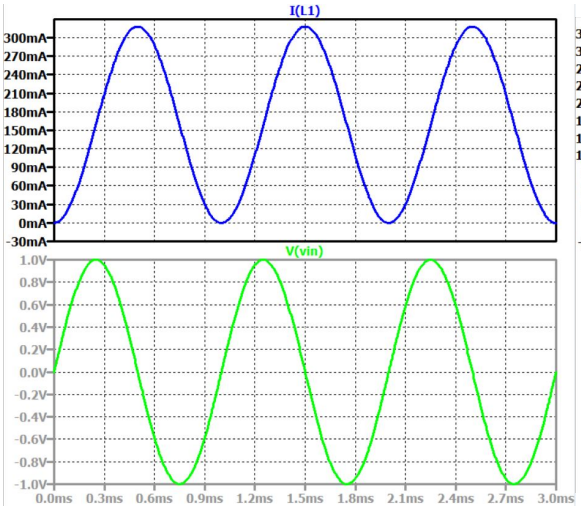
And finally, let's do an inductor:



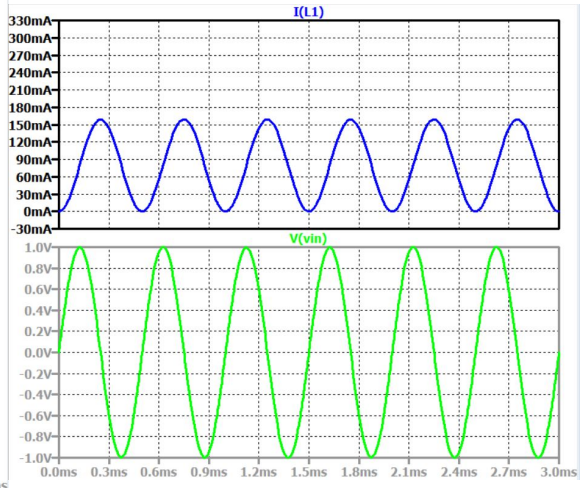
Here the current is

$$i = \frac{1}{L}\int V_{in} dt$$

$$i = -\frac{1}{L \cdot 2\pi f}\cos(2\pi f \cdot t)$$



(a) Frequency = 1kHz, I_{max} = 160mA



(b) Frequency = 2kHz, I_{max} = 80mA

Figure 7.6: V and I on a 1mH inductor. The graphs of the current contain a DC term that is not part of the sine wave response, so the current should actually be alternating positive and negative, and will be fixed in a future version.

Just like the capacitor, the frequency of the current is the same as the voltage signal, but the amplitude varies depending on the frequency. An inductor behaves like the dual of a capacitor: presenting a small resistance at low frequencies (higher current for same voltage amplitude), and a large resistance at high frequencies (lower current for same voltage amplitude).

7.3.4 Impedance of R, L, C

Since the frequency of the current is always the same as the frequency of the voltage, we can still define the voltage to current ratio, even for capacitors and inductors. In equations, impedance is represented as "Z", and like resistance it is measured in Ohms. By definition:

$$Z = \frac{V}{I}$$

For a resistor, the impedance is equivalent to the resistance, therefore:

$$Z_R = \frac{V}{I} = R$$

We can also develop equations which describe the impedance of capacitors and inductors. **The phase shift of a signal will be represented by the symbol j**. This phase shift occurs because the current through a capacitor is cos() if the voltage is sin(), since the current is the derivative of the voltage. We represent this 90 degree phase shift by adding 'j' to the resulting amplitude. We use this 'j' to remind us when we add terms, that some of the terms are sine waves, and some are cosine waves. Why we use 'j' to represent a phase shift is explained in bonus material (not needed for Engr 40M) at the end of this chapter.

From our observations, we know that for $V_{in} = 1 \cdot \sin(2\pi f \cdot t)$, the current through the capacitor should be:

$$i_{capacitor} = C \cdot 2\pi f \cos(2\pi f \cdot t)$$

$$Z_{capacitor} = \frac{V_{in}}{i_{capacitor}} = \frac{1}{j2\pi fC}$$

Notice that since the current was a cosine, we multiplied the amplitude by 'j.' We can use the same approach to compute the impedance of an inductor:

$$i_{inductor} = -\frac{1}{L \cdot 2\pi f} \cos(2\pi f \cdot t)$$

$$Z_{inductor} = \frac{V_{in}}{i_{inductor}} = \frac{1}{\frac{1}{L \cdot j2\pi f}} = j2\pi f \cdot L$$

These are general equations which hold true for describing the impedance of capacitors and inductors over all frequencies.

Example: As an exercise, now use these equations to calculate the impedance of a $1\mu F$ capacitor at 1kHz. Using this, we can calculate the current we expect to be going through the capacitor.

$$Z_c = \frac{1}{j2\pi \cdot 1kHz \cdot 1\mu F} = 159.15\Omega$$

$$I_c = \frac{1}{j159.15} = 6.3mA$$

Notice this matches the value shown on Figure 7.5a. Next let us calculate the current flowing through this same capacitor at 100kHz. It increases, as we expect.

$$Z_c = \frac{1}{2\pi \cdot 100kHz \cdot 1\mu F} = 1.59\Omega$$

$$I_c = \frac{1}{159.15} = 630mA$$

Now do repeat the above two exercises for the inductor example - L=1mH, at 1kHz. Check it against Figure 7.6a. Now calculate the current at 100kHz. Does the current increase or decrease with frequency? Is it what you expect? [a]

Questions:

Based on these equations, what can you say about the impedance of a capacitor at DC (Frequency = 0)? What can you say about the impedance of a capacitor at very high frequencies (Frequency = ∞) ? Do they look like short circuits or open circuits?

What can you say about the impedance of an inductor at DC? How about at very high frequencies?

Refer to [b] to check your answers.

[a] Current = 1.6mA, it has decreased with frequency. This is expected since the impedance of an inductor increases with frequency.

[b] A capacitor looks like an open circuit at DC, and a short circuit at very high frequencies. An inductor looks like a short circuit at DC, and an open circuit at very high frequencies.

7.3.5 Summary

The impedances of resistors, capacitors and inductors can be described by the following equations. Since the term $2\pi f$ appears so often, it is often represented simply as *omega*. You might also often see s, which represents $j2\pi f$.

$$Z_{resistor} = R$$

$$Z_{capacitor} = \frac{1}{j2\pi fC} = \frac{1}{j\omega C} = \frac{1}{sC}$$

$$Z_{inductor} = j2\pi fL = j\omega L = sL$$

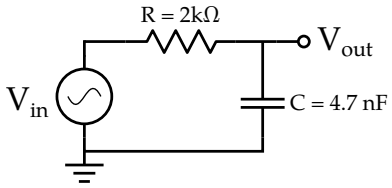
$$\omega = 2\pi f, \quad s = j2\pi f$$

By using impedance, we can treat capacitors and inductors like resistors when analyzing them in the circuit, but the resulting circuit behavior is now frequency dependent.

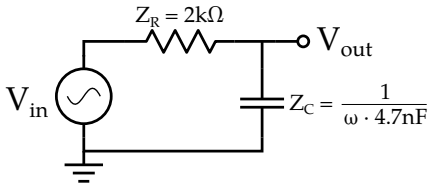
7.4 Filters - Transfer Functions and Bode Plots

Knowing that adding capacitors and inductors create frequency dependent circuits, we can now explore ways of using them to do more interesting things with our circuits. By using the idea of impedance, we can analyze the circuit using the tools we developed to analyze resistor circuits in the first part of the class. You can use nodal analysis, series/parallel reduction, voltage/current dividers, etc.

Let us use this new method to analyse the RC circuit below, to understand how its behavior depends on frequency. That is, for any sine wave that we put in, we want to find the amplitude of the corresponding output. This is known as the *frequency response*, or sometimes as the *transfer function*.



The impedance of a capacitor is $Z_C = \frac{1}{j\omega C}$. The impedance of a resistor is simply $Z = R$.



Now that the circuit is expressed in terms of impedance, the output voltage is just the result of a voltage divider:

$$V_{out} = V_{in} \cdot \frac{Z_C}{Z_R + Z_C}$$

The ratio between the output and input voltages, known as the *gain*, is therefore just

$$\text{Gain} = \frac{V_{out}}{V_{in}} = \frac{Z_C}{Z_R + Z_C}$$

Plugging in the values for this circuit, we can write

$$\text{Gain} = \frac{\frac{1}{j\omega \cdot 4.7\,\text{nF}}}{2\,\text{k}\Omega + \frac{1}{j\omega \cdot 4.7\,\text{nF}}}$$

And multiplying through by $s \cdot 4.7\,\text{nF}$ gives

$$\text{Gain} = \frac{1}{j\omega \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF} + 1}$$

It's worth making a couple observations at this point. First, the gain will never be more than 1. This makes sense, because the output is the result of a voltage *divider* and must be some fraction of the input. Second, the gain decreases as frequency ($\omega = 2\pi f$) increases. In other words, as the frequency increases, the capacitor impedance decreases, and the output amplitude becomes less and less.

In other words, this circuit behaves like a *low pass filter*. When a set of tones of varying frequencies but the same amplitude are fed into this circuit, the low frequency tones appear at V_{out} at the same amplitude as they were at the input, but the high frequency tones appear at V_{out} at lower amplitudes (in other words, the amplitude of the signal decreases as frequency increases).

Now that we have a transfer function describing the gain of this circuit, let's plot it on a Bode plot, which is what we use to represent these gain-frequency relationships.

7.4.1 Plotting the Transfer Function

First, let's use a brute-force approach to plotting, and then we'll work backward to the intuition, and then work out a quick way to plot the frequency response without a computer (and discover why Bode plots are so useful).

Gain is usually expressed in decibels (dB), so we need to convert our gain equation to dB. Remember that $\text{Gain}_{\text{dB}} = 20 \cdot \log_{10}(V)$.

Pull out your favorite plotting tool and plot the gain, using a logarithmic X scale for frequency and a linear Y scale for dB (since dB is already a log scale). Python/NumPy and MATLAB examples are at the end of this document.

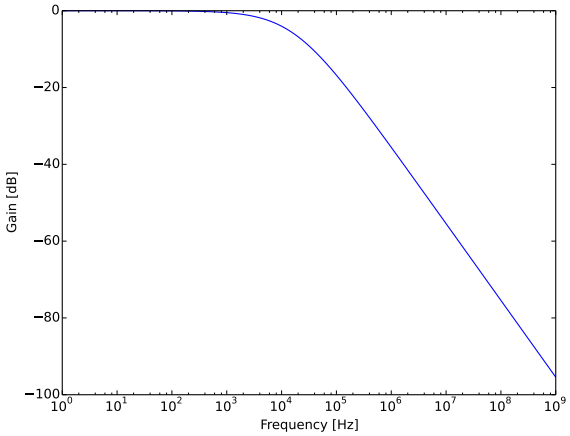


Figure 7.7: Bode plot of simple RC circuit

The gain basically has two straight lines, connected by a smooth curve. The first line is at 0 dB and goes from DC (0 Hz) to about 1 kHz. The second line is a downward slope where the gain steadily drops as frequency increases. The point at which these two straight lines would intersect is another important feature of the Bode plot. It is a corner, and therefore we call it a *corner frequency*. Bode plots can have multiple corner frequencies - in the case of our example above, there is only one.

When drawing a Bode plot, the first step we usually take is to find this corner frequency. To do this we first write down the gain equation:

$$\text{Gain}_{\text{dB}} = 20 \cdot \log_{10}(|\frac{1}{j2\pi f \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF} + 1}|)$$

Looking at this equation we can see where the two "lines" come from. For low frequencies, the "j term" in the denominator is much smaller than 1, and can be ignored (remember this is a log/log plot so each step in the x direction is a 10x change in frequency). Similarly, at high frequencies, the "j term" in the denominator is much larger than the 1, and the one can be ignored. The corner frequency is when these two approximations have the same magnitude response, or when:

$$\text{Gain}_{\text{dB}} = 20 \cdot \log_{10}(\frac{1}{1}) = 20 \cdot \log_{10}(\frac{1}{2\pi f \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF}})$$

Setting the two terms equal to each other (since the corner frequency is the point when the two curves have the same gain),

$$2\pi f_c \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF} = 1$$

$$f_c = \frac{1}{2\pi \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF}} = 16.9\,\text{kHz}$$

We can now see the role that the corner frequency, f_c, plays. For frequencies *less* than f_c, we can make the approximation that the $j2\pi f$ term is much smaller than 1, resulting in the flat line we draw to the left of the corner frequency. And for the frequencies *more* than f_c, we can make the approximation that the $j2\pi f$ term is larger than 1, resulting in the sloped line that we draw to the right of the corner frequency in this example.

We can find the slope of the line after the corner frequency mathematically - and in fact you will be often asked to do so for non-zero slopes, as this is an important feature of a Bode plot. At higher frequencies, the first term ($j2\pi f \cdot 2k\Omega \cdot 4.7nF$) is large. It will be much larger than 1, and we can treat the +1 as negligible and simply drop it, to write the following:

$$\text{Gain}_{\text{dB}} = 20 \cdot \log_{10}(\frac{1}{2\pi f \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF}})$$

$$\text{Gain}_{\text{dB}} = 20 \cdot (\log_{10}(1) - \log_{10}(2\pi f \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF}))$$

$$\text{Gain}_{\text{dB}} = 0 - 20 \cdot \log_{10}(2\pi \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF}) - 20 \cdot \log_{10}(f)$$

$$\text{Gain}_{\text{dB}} = -20 \cdot \log_{10}(f) - 20 \cdot \log_{10}(2\pi \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF})$$

All the terms in the above equation are constants except for$-20 \cdot \log_{10}(f)$, indicating that the slope is a decrease by 20dB for every 10x increase in frequency. Notice that the slope of this line doesn't depend on the component values. You may often hear electrical engineers referring to this as "20dB per decade."

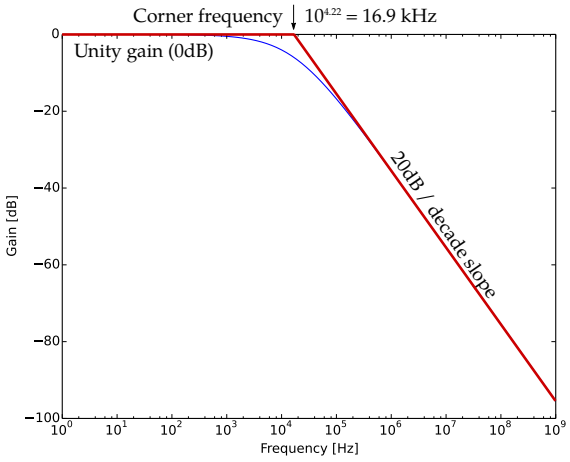
This result points out one of the important features of Bode (log-log) plots: any response region that is proportional to f^n is a straight line, with a slope of $20 \cdot n$ dB/decade, for any value of n, positive or negative. The fact that these regions are straight lines on a log-log plot makes it easier to plot, and find the corner frequencies.

Using the corner frequency, also provides a simplified way to write the gain equation:

$$\text{Gain}_{\text{dB}} = 20 \cdot \log_{10}(\frac{1}{2\pi f \cdot 2\,\text{k}\Omega \cdot 4.7\,\text{nF} + 1}) = 20 \cdot \log_{10}(\frac{1}{\frac{f}{f_c} + 1}) = 20 \cdot \log_{10}(\frac{1}{\frac{\omega}{\omega_c} + 1})$$

where $\omega_c = \frac{1}{2\,\text{k}\Omega \cdot 4.7\,\text{nF}} = 2\pi f_c$. Below f_c (or ω_c if you are using radians) the gain will be constant, and above f_c the gain falls at 20dB per decade of frequency.

Using these three bits of information: the corner frequency, the gain below the corner frequency, and the gain above the corner frequency, we can draw an approximate Bode plot representation as shown in the figure below in pink. You'll notice this approximation very closely follows the plot which was drawn using computer software, and was actually fairly simple to create.



**An alternative way of seeing the 20dB/dec slope is to plot a few points on the graph for yourself. Choose points which are a decade apart - eg. let's choose points $\omega = 10\omega_c, 100\omega_c, 1000\omega_c$.

$$Gain_{dB}(\omega) = 20 \cdot \log_{10}(\frac{1}{1+\frac{\omega}{\omega_c}})$$

$$Gain_{dB}(10\omega_c) = 20 \cdot \log_{10}(\frac{1}{1+\frac{10\omega_c}{\omega_c}}) = 20 \cdot \log_{10}(\frac{1}{1+10}) \approx 20 \cdot log(0.1) = -20dB$$

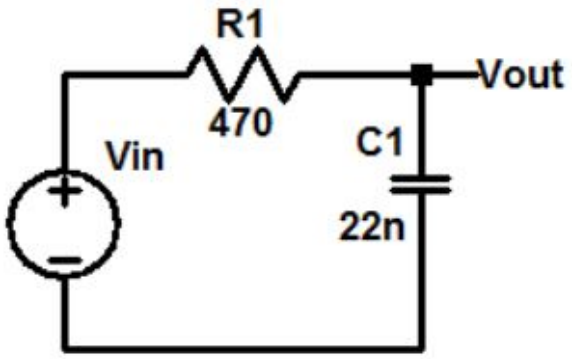
$$Gain_{dB}(100\omega_c) = 20 \cdot \log_{10}(\frac{1}{1+\frac{100\omega_c}{\omega_c}}) = 20 \cdot \log_{10}(\frac{1}{1+100}) \approx 20 \cdot log(0.01) = -40dB$$

$$Gain_{dB}(100\omega_c) = 20 \cdot \log_{10}(\frac{1}{1+\frac{1000\omega_c}{\omega_c}}) = 20 \cdot \log_{10}(\frac{1}{1+1000}) \approx 20 \cdot log(0.001) = -60dB$$

This information is summarized in Table 7.1, and it becomes obvious that after the corner frequency, for every decade increase in frequency, the gain decreases by 20dB. If you were to plot these on a graph, you would end up with a straight line of -20dB/dec slope!

An Example

Find the transfer function of the following circuit, and its corner frequency. Plot the transfer function of this circuit on a Bode plot, indicating its corner frequency, and the value of any slopes.



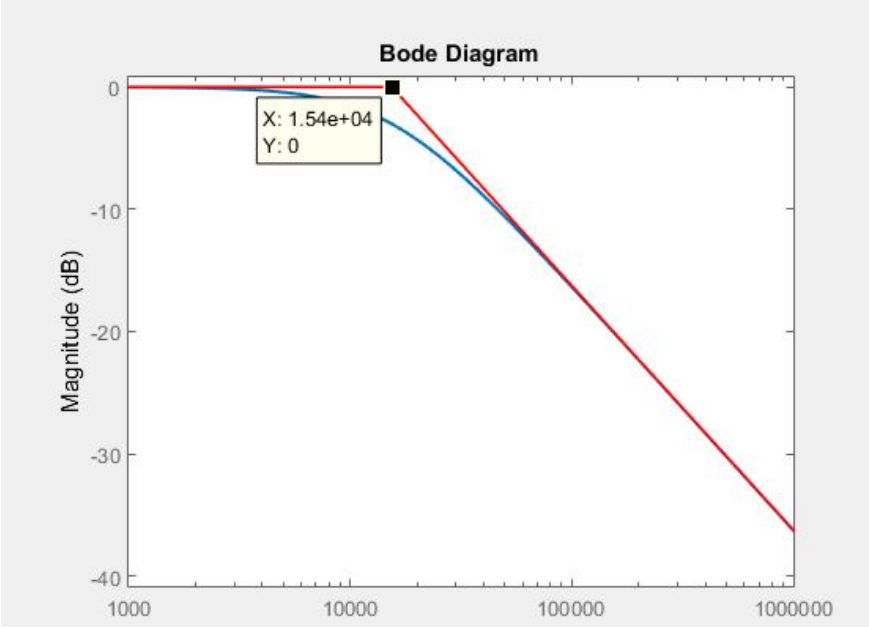
$$Z_R = R = 470\Omega$$
$$Z_C = \frac{1}{j\omega C} = \frac{1}{j\omega \cdot 22nF}$$
$$Gain = \frac{V_{out}}{V_{in}} = \frac{Z_C}{Z_R + Z_C} = \frac{\frac{1}{j\omega \cdot 22nF}}{470\Omega + \frac{1}{j\omega \cdot 22nF}} = \frac{1}{1 + j\omega \cdot 22nF \cdot 470\Omega}$$
$$\text{Rewrite as: } Gain = \frac{1}{1 + \frac{j\omega}{\omega_c}} \quad where \quad \omega_c = \frac{1}{22nF \cdot 470\Omega}$$
$$\text{Therefore: } f_c = \frac{1}{2\pi \cdot 22nF \cdot 470\Omega} = 15.4kHz$$

Now that we know the corner frequency, we need to find the slopes of the two lines that meet at the corner frequency.

$$Gain_{dB} = 20 \cdot \log_{10}(\frac{1}{1 + \frac{\omega}{\omega_c}})$$

When $\omega << \omega_c$, then $\frac{\omega}{\omega_c} << 1$ and $Gain_{dB} \approx 20 \cdot \log_{10}(\frac{1}{1}) = 0dB$ so we draw a straight line of 0dB gain up to the corner frequency. When $\omega >> \omega_c$, then $\frac{\omega}{\omega_c} >> 1$ and $Gain_{dB} \approx 20 \cdot \log_{10}(\frac{1}{\frac{\omega}{\omega_c}}) = 20 \cdot \log_{10}\omega_c - 20 \cdot \log_{10}(\omega)$ we draw a straight line of -20dB/dec above the corner frequency. These two lines intersect at the corner frequency.

The plot can be drawn simply using asymptotes as indicated in red on the following figure, as well as the actual output in blue:



7.4.2 Dealing with the Phase Shift

Now let's look at what happens when the circuit operates near the corner frequency. At first it seems like the gain at the corner frequency should be one half, since the two terms in the denominator are both around 1. But the actual value is 0.707, $1/\sqrt{2}$ and not 1/2. The reason for this is that we are adding two terms, but one is phase shifted from the other. That is, we are not adding together two sine waves of the same magnitude. Instead we are adding together a sine wave with a cosine wave of the same magnitude.

To understand what the results should be, lets look at what type of sine wave is created by $1 + jx$ (at f_c this becomes $1 + j$). This sum really means $sin(\omega t) + x \cdot cos(\omega t)$, which means we are trying to find the magnitude, A, and phase ϕ such that:

$$A \cdot sin(\omega t + \phi) = sin(\omega t) + x \cdot cos(\omega t)$$

Since this equation needs to be true for all values of t, we can choose values of t which make finding A and ϕ easier. We will choose $t = 0$, to make the sin term zero, and $t = \frac{\pi}{2 \cdot \omega}$ to make the cos term zero. This yields the following equations:

$$A \cdot sin(\phi) = sin(0) + x \cdot cos(\omega \cdot 0) = x$$

$$A \cdot sin(\pi/2 + \phi) = A \cdot cos(\phi) = sin(\pi/2) + x \cdot cos(\pi/2) = 1$$

Since $A \cdot sin(\phi)$ is x and $A \cdot cos(\phi)$ is 1, and we know that $sin(\phi)^2 + cos(\phi)^2 = 1$, $A = \sqrt{1^2 + x^2}$. Another way of seeing the same result is to realize that 'A' is the hypotenuse of a right triangle with the other sides being 1 and x. This means that the value of 'A' is the square root of the sum of the squares of the sides, which is $\sqrt{1^2 + x^2}$.

Table 7.1: Plotting a Bode plot using pointss

$\omega(rad/s)$	$Frequency(Hz)$	$Gain_{db}(dB)$
$10\omega_c$	$169kHz$	$-20dB$
$100\omega_c$	$1.69MHz$	$-40dB$
$1000\omega_c$	$16.9MHz$	$-60dB$

7.4.3 More RC circuits

Let's consider another RC filter, this time configured a little differently. We want to find its frequency response/transfer function, and plot this on a Bode plot.

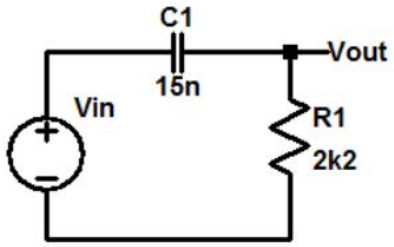


Figure 7.8: Another RC circuit

First, consider this circuit qualitatively. At low frequencies, the capacitor presents a very large resistance. Hence we expect $\frac{V_{out}}{V_{in}}$ to be small at low frequencies. Conversely, at high frequencies the capacitor presents a very small resistance. Therefore we expect $\frac{V_{out}}{V_{in}}$ to be high (approaching unity as the capacitor approaches becoming a short circuit).

Now let's look at this quantitatively and see if the results match what we expect. Start by finding the impedance of the capacitor and the resistor.

$$Z_R = R = 2.2k\Omega$$

$$Z_C = \frac{1}{\omega C} = \frac{1}{\omega 15nF}$$

$$\frac{V_{out}}{V_{in}} = \frac{Z_R}{Z_R + Z_C} = \frac{2.2k\Omega}{2.2k\Omega + \frac{1}{\omega \cdot 15nF}}$$

$$Gain = \frac{V_{out}}{V_{in}} = \frac{\omega \cdot 15nF \cdot 2.2k\Omega}{\omega \cdot 15nF \cdot 2.2k\Omega + 1}$$

By simply looking at the above equation, can you see that the gain approaches zero at low frequencies and unity at high frequencies? Let's rewrite the equation so that it's easier to see this.

$$Gain = \frac{V_{out}}{V_{in}} = \frac{\frac{\omega}{\omega_c}}{\frac{\omega}{\omega_c} + 1} \qquad \text{where} \qquad \omega_c = \frac{1}{RC} = \frac{1}{2.2k\Omega \cdot 15nF}$$

Therefore the corner frequency, $f_c = \frac{1}{2\pi \cdot 2.2k\Omega \cdot 15nF} = 4.8kHz$

When $\omega >> \omega_c$, then $\frac{\omega}{\omega_c} >> 1$ and the gain will be about 1, or 0dB, as we expected. If $\omega << \omega_c$, then $\frac{\omega}{\omega_c} << 1$ and:

$$Gain_{dB} \approx 20 \cdot \log_{10}(\frac{\omega}{\omega_c}) = 20 \cdot \log_{10}(\omega) - 20 \cdot \log_{10}(\omega_c)$$

Below the corner frequency, the gain increases at 20dB/dec up to the value of the gain at the corner frequency. Using this information, we can draw two asymptotes crossing at the corner frequency to represent the Bode plot of this circuit. This is shown in Figure 7.9.

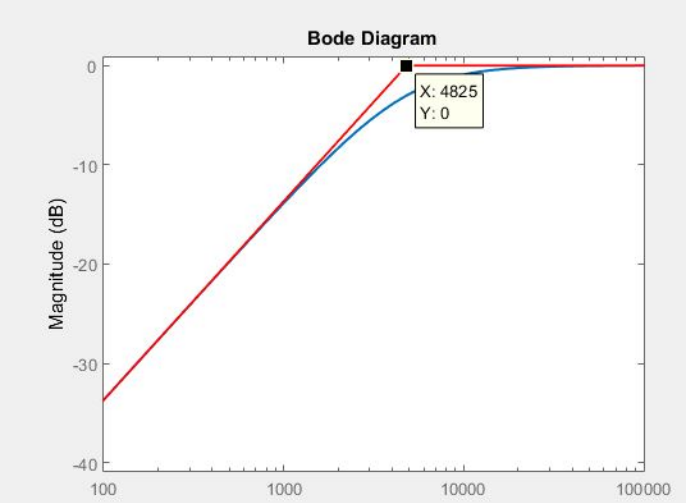


Figure 7.9: Simple Bode plot of the high pass filter

By swapping the capacitor and the resistor around, we have created a high pass filter (low gain at low frequencies, and unity gain at higher frequencies). The actual response is shown in blue in that plot, and the only significant deviation is at the corner frequency (which has a gain of -3dB rather than 0dB).

Finally, let's look at a more complex RC circuit. This circuit has two capacitors.

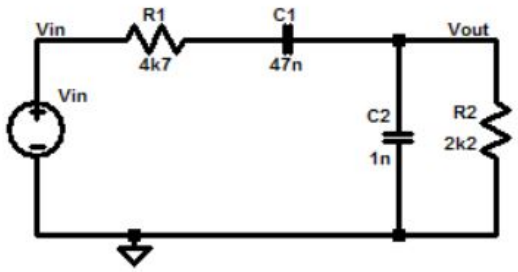


Figure 7.10: A slightly more complex RC circuit

$$Z_R1 = R1 = 4.7k\Omega$$
$$Z_R2 = R2 = 2.2k\Omega$$
$$Z_C1 = \frac{1}{\omega C1} = \frac{1}{\omega \cdot 47nF}$$
$$Z_C2 = \frac{1}{\omega C2} = \frac{1}{\omega \cdot 1nF}$$

Let's call the series combination of R2 and C2 a lump impedance Z2, and the parallel combination of R1 and C1 a lump impedance Z1, as shown in Figure 7.11.

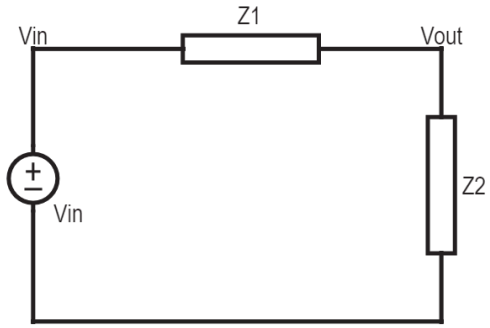


Figure 7.11: Abstracting the circuit to obtain the transfer function

Then the circuit is simply a voltage divider:

$$Gain = \frac{V_{out}}{V_{in}} = \frac{Z2}{Z2 + Z1}$$

Writing down the actual impedance for Z1 and Z2:

$$Z1 = Z_{R1} + Z_{C1} = R1 + \frac{1}{\omega \cdot C1} = \frac{1 + \omega \cdot R1 \cdot C1}{\omega \cdot C1}$$

$$Z2 = Z_{R2} \parallel Z_{C2} = \frac{1}{\frac{1}{R2} + \omega \cdot C2} = \frac{R2}{1 + \omega \cdot R2 \cdot C2}$$

Substituting these back into the gain equation we get:

$$Gain = \frac{\frac{R2}{1+\omega \cdot R2 \cdot C2}}{\frac{1+\omega \cdot R1 \cdot C1}{\omega \cdot C1} + \frac{R2}{1+\omega \cdot R2 \cdot C2}}$$

While this looks bad, it will look much better after a little algebra. Multiply numerator and denominator by $(\omega C1) \cdot (1 + \omega \cdot R2 \cdot C2)$ to to get rid of the fractions on the bottom:

$$= \frac{\omega \cdot R2 \cdot C1}{(1+\omega \cdot R1 \cdot C1) \cdot (1+\omega \cdot R2 \cdot C2) + \omega \cdot R2 \cdot C1}$$

While this is a little more complex than the previous gain formulas, we can still find the lines that make up the gain plot, by systematically looking at the equation at in different frequency bands separated by the frequencies[2] we identify: $\omega_1 = \frac{1}{R2 \cdot C1} = 9.7krad/s = 1.5kHz$ and $\omega_2 = \frac{1}{R1 \cdot C2} = 213krad/s = 34kHz$.

- At low frequencies: When F (ω) is very small, the denominator will be about 1 ($1 >> \omega(R1 \cdot C1 + R2 \cdot C2 + R2 \cdot C1)$, so the gain will be $\omega \cdot R2 \cdot C1 = \omega/\omega_1$. This means at low frequencies, we will draw a line increasing at 20dB/dec, intersecting with the 0dB line at frequency ω_1
- Next, we look at what happens when the frequency is large enough that ($\omega(R1 \cdot C1 + R2 \cdot C2 + R2 \cdot C1 >> 1$) but below ω_2 so the ω^2 term is still small. Then the equation would be approximated by the following:

$$\approx \frac{\omega \cdot R2 \cdot C1}{(\omega \cdot R1 \cdot C1) \cdot (1) + \omega \cdot R2 \cdot C1}$$
$$= \frac{\omega R2}{\omega R1 + \omega R2} = \frac{R2}{R1 + R2} = -9.9dB$$

  It is just a resistive divider! Hence, in this range, the circuit is simply a flat line (slope = 0) at a magnitude of -9.9dB.

- Finally, we consider what happens when the frequency is high. Now the denominator will be dominated by the ω^2 term:

$$\approx \frac{\omega \cdot R2 \cdot C1}{(\omega \cdot R1 \cdot C1) \cdot (\omega \cdot R2 \cdot C2)}$$
$$= \frac{1}{\omega \cdot R1 \cdot C2} = \frac{\omega_2}{\omega}$$

This means at high frequencies (above ω_2) we will draw a line decreasing at -20dB/dec.

[2]Notice that these frequencies will not actually be the corner frequency of the curve, since these are the frequencies where the gain would become one if this term dominates. As you will find out shortly, this function never becomes one, so the intersection points (the actual corner frequencies) are not these constants

The final Bode plot would look like the following. It is a bandpass filter.

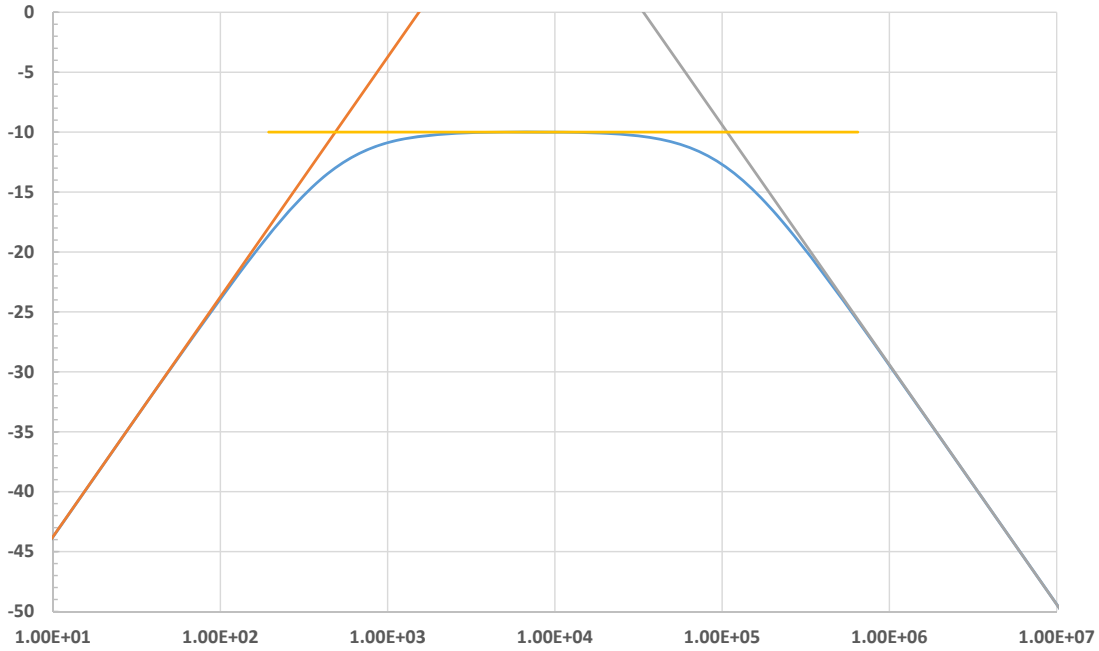


Figure 7.12: Bode plot of bandpass filter. The actual filter response is shown in blue. Also shown is the low frequency approximation in orange, middle frequency in yellow, and high frequency in grey.

Problem 7.1 Derive the gain of this circuit and sketch the Bode plot of the frequency response. Use what you've learned about capacitor behaviour at high and low frequencies to check that your answers make sense.

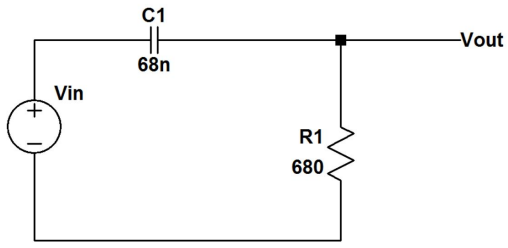


Figure 7.13: RC circuit example 1

Problem 7.2 Derive the gain of this circuit and sketch the Bode plot of the frequency response. Use what you've learned about capacitor behaviour at high and low frequencies to check that your answers make sense. This question is a little different from the problems we have done before, since the max gain is not one.

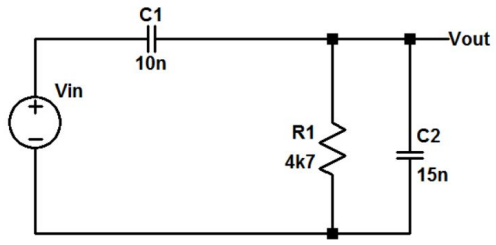


Figure 7.14: RC circuit example 2

7.4.4 Summary

You have now seen how to create and analyze a low pass filter, high pass filter, and bandpass filter.

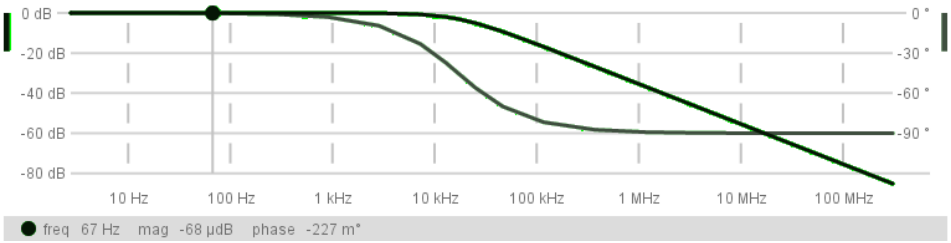
To summarize, when you are given an RLC circuit and asked to find the frequency response:

1. Find the impedance for each element in the circuit.
2. Solve for the output in terms of the input to get the gain.
3. Convert to dB and plot.
4. The resulting plot can be approximated by a set of straight lines where it is easy to estimate one point of the line, and its slope. If the gain is changing with F the slope with be 20dB/decade, and if it is changing by 1/F it is -20dB/decade. If the gain is changing by F^2, then it will be 40dB/decade (since in log scale, squaring something just multiplies it by 2).
5. This straight line approximation will provide the all the information you need, including the corner frequencies. The position of the lines and the corner frequencies does not change when you use the "correct" formulas that incorporate phase information.
6. Of course if you want an exact plot you should use a computer.

7.5 Using EveryCircuit

EveryCircuit can do a frequency domain simulation, which is an excellent way to check your answers or gain intuition about how the circuits work.

Build your circuit, and view the voltages at both the input and output nodes (click on the node, then click the "eye" in the lower-left corner). Then click the yellow "Run AC" button, labeled with an 'f'. The bright green line is the gain; the pale green line is the phase. You can zoom and pan the plot if it doesn't show the frequency range you're interested in.



7.6 BONUS MATERIAL - Why Use j for Phase

In this chapter we have used the letter 'j' to remind us that this term is really phase shifted from the input; it is a cosine wave, and not a sine wave. We used this notation since there is a very useful relationship between sine and cosine waveforms, and imaginary exponentials. The definition of an exponential function is that the derivative of the function is equal to the function times a constant, which is why the output of an RC circuit decays with an exponential waveform:

$$\frac{d}{dt}e^{st} = s \cdot e^{st}$$

If we take the derivative again, we will get:

$$\frac{d^2}{dt^2}e^{st} = s^2 \cdot e^{st}$$

Now let's look at what happens when we do this with a sine wave:

$$\frac{d}{dt}\sin(\omega t) = \omega \cdot \cos(\omega t)$$

Taking the derivative again gives:

$$\frac{d^2}{dt^2}\sin(\omega t) = -\omega^2 \cdot \sin(\omega t)$$

Notice that if we look at the second derivative lines, sin() and exp() don't look that different. Both functions are unchanged after the second derivative, and both are multiplied by a constant squared. Of course the first derivatives are different, and the second derivative of sin() has this nasty negative sign. To make them more similar, I need to make $s^2 = -\omega^2$. Of course that is impossible if we are dealing with real numbers, but it is easy to do if I can use imaginary numbers. Imaginary numbers are numbers that when squared are negative and are typically written as $x \cdot j$ where $j = \sqrt{-1}$.[3] So I can make the two functions look more similar if I make $s = j \cdot \omega$. But what does an exponential with an imaginary time constant mean?

To help work some of this out, let's define two functions. The first function, g, is going to an exponential, which gives the expected result when you take its derivative:

$$g(t) = e^{st} \qquad \frac{d}{dt}g(t) = s \cdot e^{st} = s \cdot g(t)$$

[3]Electrical engineers use i to represent current so they typically use j to represent $\sqrt{-1}$, while the rest of the world use $i = \sqrt{-1}$

So far no surprises. Now let me define another function h. This function returns a complex number (it has a real part and an imaginary part), and is the sum of a cosine wave and an imaginary sine wave. Taking the derivative of this function is also easy:

$$h(t) = \cos(\omega t) + j \cdot \sin(\omega t)$$
$$\frac{d}{dt}h(t) = \omega \cdot [-sin(\omega t) + j \cdot cos(\omega t)]$$

Now here comes the surprising part. If you look at the derivative of h(t), it turns out to be a constant time h(t). The cos term is multiplied by $j\omega$, and the sin term, was also multiplied by $j\omega$ making it now real and negative. In other words:

$$\frac{d}{dt}h(t) = i\omega \cdot h(t)$$

Notice that this equation is exactly the same equation as the equation for g, if $s = j \cdot \omega$. Said differently, we just figured out what a complex exponential represents:

$$e^{j\omega t} = \cos(\omega t) + j \cdot \sin(\omega t)$$

This is a very famous result in mathematics and is known as Euler's equation! Since cos(-x) = cos (x), and sin(-x) = -sin(x), we have:

$$e^{i\omega t} = cos(\omega t) + j \cdot sin(\omega t); \qquad e^{-i\omega t} = cos(\omega t) - j \cdot sin(\omega t)$$

$$\sin \omega t = \frac{e^{j\omega t} - e^{-j\omega t}}{2j}; \qquad \cos \omega t = \frac{e^{j\omega t} + e^{-j\omega t}}{2}$$

Thus sinusoidal inputs are really just the sum of two exponential function (with complex time constants), and since the system is linear, we can look at the response to each exponential individually. This is great, since we know the derivative of an exp is always an exponential, and we don't have to worry about sin and cosine waves. So the precise definition of impedance is simply:

$$Z_R = \frac{V_{in}}{i_R} = R$$
$$Z_C = \frac{V_{in}}{i_C} = \frac{1}{sC}$$
$$Z_L = \frac{V_{in}}{i_L} = sL$$
$$\text{where } s = j\omega = j \cdot 2\pi f$$

So if we use a complex exponential to drive the circuit, we are driving it with a cos() + j sin() waveform. Notice if I multiply this waveform by 'j', I get waveform with a -sin() as the real part, which is a $\pi/2$ phase shift from my original waveform. So when working with complex exponentials, the time constant is really imaginary, $(j \cdot \omega)$, and multiplying that function by 'j' does really change the phase of the waveform. So if you take more circuits or signal processing classes, you will see complex exponentials, and 'j' again.

7.7 Solutions to Practice Examples

Solution 7.1:

$$Z_R = R = 680\Omega$$

$$Z_C = \frac{1}{\omega C} = \frac{1}{\omega 68nF}$$

$$\frac{V_{out}}{V_{in}} = \frac{Z_R}{Z_R + Z_C} = \frac{680\Omega}{680\Omega + \frac{1}{\omega \cdot 68nF}}$$

$$Gain = \frac{V_{out}}{V_{in}} = \frac{\omega \cdot 68nF \cdot 680\Omega}{\omega \cdot 68nF \cdot 680\Omega + 1} = \frac{\frac{\omega}{\omega_c}}{\frac{\omega}{\omega_c} + 1}$$

$$\text{Therefore } \omega_c = \frac{1}{68nF \cdot 680\Omega}$$

Therefore the corner frequency $f_c = \frac{\omega_c}{2\pi} = \frac{1}{2\pi \cdot 68nF \cdot 680\Omega} = 3.44kHz$. From the gain equation we can see that the gain approaches zero at low frequencies, and approaches unity at high frequencies. Using this information, we can construct a simple Bode plot using straight lines as shown.

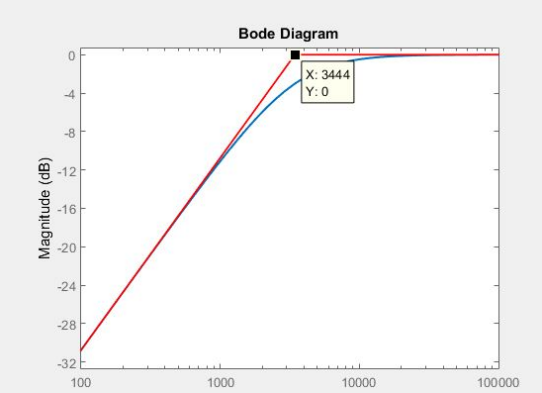


Figure 7.15: RC circuit example 1 - Simplified hand drawn Bode plot

An accurate Bode plot is shown in Figure 7.16.

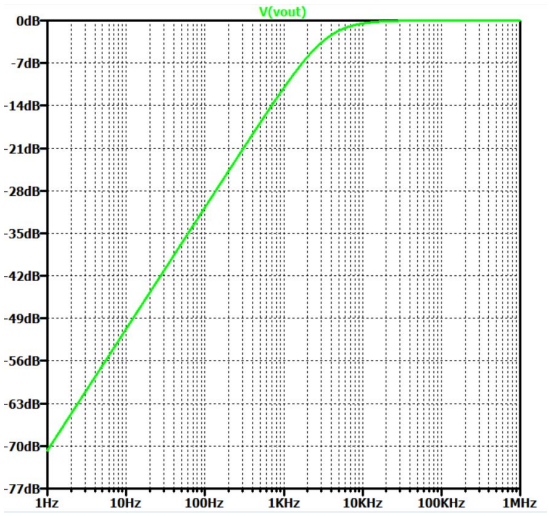


Figure 7.16: RC circuit example 1 - Bode plot

Solution 7.2:
If we just think about the circuit qualitatively, it be thought of as a voltage divider comprising two elements - Z_{C1} and $Z_R \parallel Z_{C2}$. In other words:

$$\frac{V_{out}}{V_{in}} = \frac{Z_R \parallel Z_{C2}}{Z_{C1} + Z_R \parallel Z_{C2}}$$

The parallel combination of R1 and C2 is dominated by R1 at low frequencies, since C2 appears to be an open circuit. Similarly, C1 appears as an open circuit (or a very big impedance) and therefore the fraction of Vin that appears at Vout will be very small at low frequencies.

We can then derive the transfer function and Bode plots, and check them by seeing if they match our intuition of how the circuit should behave.

$$Z_R = R = 4.7k\Omega$$
$$Z_{C1} = \frac{1}{\omega C1} = \frac{1}{\omega 10nF}$$
$$Z_{C2} = \frac{1}{\omega C2} = \frac{1}{\omega 15nF}$$
$$Z_R \parallel Z_{C2} = \frac{Z_R \cdot Z_{C2}}{Z_R + Z_{C2}} = \frac{R \cdot \frac{1}{\omega C1}}{R + \frac{1}{\omega C1}} = \frac{R}{1 + \omega \cdot R \cdot C2}$$
$$Gain = \frac{V_{out}}{V_{in}} = \frac{\frac{R}{1+\omega \cdot R \cdot C2}}{\frac{1}{\omega C1} + \frac{R}{1+\omega \cdot R \cdot C2}} = \frac{1}{\frac{1}{\omega C1} \cdot \frac{1+\omega \cdot R \cdot C2}{R}} = \frac{1}{1 + \frac{1+\omega \cdot R \cdot C2}{\omega \cdot R \cdot C1}} =$$

$$\frac{\omega \cdot R \cdot C1}{1 + \omega \cdot R \cdot C1 + \omega \cdot R \cdot C2} = \frac{\omega \cdot R \cdot C1}{1 + \omega \cdot R \cdot (C1 + C2)}$$

The numerator shows that the gain approaches zero at low frequencies. At high frequencies, the 1 becomes insignificant compared to other terms, and the equation simplifies to a capacitive divider, as we expected, ie. $Gain = \frac{C1}{C1 + C2}$.

The denominator is in a form we are familiar with, indicating that there is a corner frequency somewhere, where $\omega_c = \frac{1}{\omega \cdot R \cdot (C1 + C2)}$.

To draw the Bode plot, we need to calculate the corner frequency and the gain at high frequencies.

$$f_c = \frac{\omega_c}{2\pi} = \frac{1}{2\pi \cdot R \cdot (C1 + C2} = 1.35kHz$$

$$Gain_{dB} \text{ at high frequencies} = 20 \cdot \log \frac{C1}{C1 + C2} = 20 \cdot log0.4 = -7.96dB$$

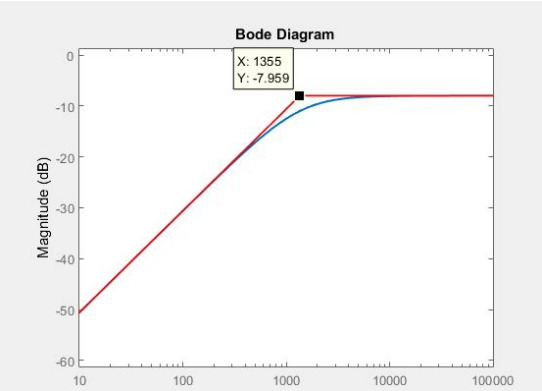


Figure 7.17: RC circuit example 2 - Simplified hand drawn Bode plot

An accurate Bode plot is shown in Figure 7.18.

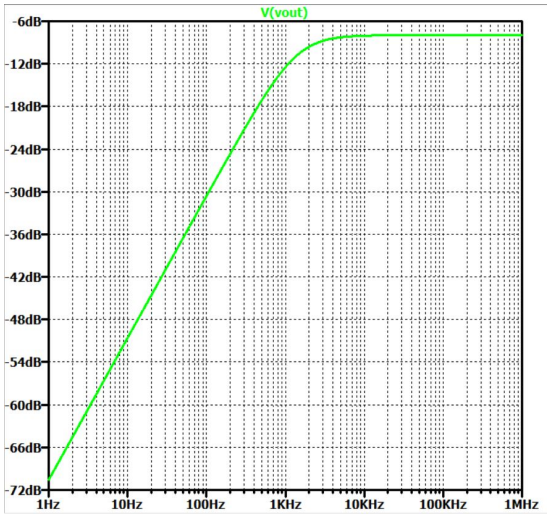


Figure 7.18: RC circuit example 2 - Accurate Bode plot

Chapter 8

Operational amplifiers

An operational amplifier is a device with two inputs and one output. It takes the difference between the voltages at the two inputs, multiplies by some very large gain, and outputs the result. Like most devices, operational amplifiers aren't useful *by themselves*—not even for amplifying voltages. Rather, we use them in circuits with resistors, capacitors and other devices, to build circuits that do intelligent and precise things.

Before we dive in, a word of advice. Operational amplifiers can *look* intimidating, because they have lots of terminals and scary statements like "infinite gain". The intuition's a little tricky at first, but the reason they're so popular is that they vastly simplify the design of circuits that manipulate and process signals. Once you get the hang of it, they're very straightforward—so hang in there, it'll pay off.

8.1 Getting to Know the Op-Amp

On schematic circuit diagrams, an *operational amplifier*, or *op-amp* for short, is represented using the symbol shown in Figure 8.2.

Terminals of an op-amp

It has five terminals:

- The *positive power supply terminal* is at the top of the symbol, and is almost always connected

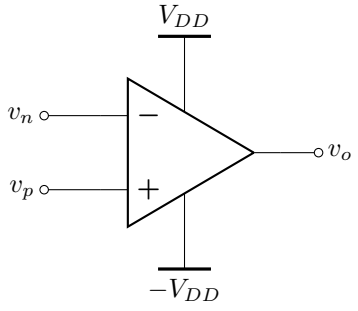


Figure 8.1: Schematic symbol for an op-amp

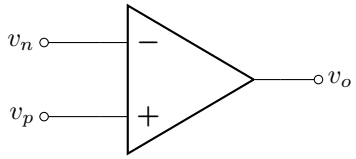


Figure 8.2: Schematic symbol for an op-amp, with power supply terminals omitted

to V_{DD}.

- The *negative power supply terminal* is at the bottom of the symbol, and is typically connected to $-V_{DD}$, a (different) negative power supply. However, precisely where it's connected is a design decision; in some circuits, including the electrocardiogram you'll build in lab 4, it is connected to ground.
- The *inverting input* is labeled − on the schematic symbol.
- The *non-inverting input* is labeled + on the schematic symbol.
- The fifth terminal, of course, is the output.

Omitting power supply terminals

In practice, we often know implicitly what the power supplies of the circuit are. In these situations, it's common to omit the power supply terminals from the symbol, to make the circuit diagram less cluttered. Such a symbol is shown in Figure 8.2. Although two of the terminals aren't drawn, they're still there! Don't forget to connect them in your circuit.

Another common gotcha: There's no convention about whether the schematic symbol should be drawn with the inverting (−) input or non-inverting (+) input on top. In this class, we'll often draw the inverting input on top, because it's often more convenient this way. But sometimes the reverse is easier, and it's on you to check the sign with which the input is labeled.

Notation of voltages

People use lots of different notations for the input, output and power voltages of an op-amp. In this class, we'll use v_p to mean the non-inverting input voltage, v_n to mean the inverting input voltage, and v_o to mean the output voltage. Some texts use v^+ and v^-, or v_+ and v_-, instead of v_p and v_n. Confusingly, different texts and datasheets also use V_+ and V_- to mean either the inputs or the power supply terminals. For this reason, we'll avoid this notation, but watch out for it in datasheets and other sources.

Behavior of op-amp

Now, what does an op-amp *do*? An op-amp takes the difference between its two inputs, $v_p - v_n$, multiplies it by an absurdly large number A, and its output voltage v_o is the result. So the behavior of an op-amp is described by the equation

$$v_o = A(v_p - v_n). \tag{8.1}$$

Gain of op-amp

The number A is called the *gain* of the op-amp, and it is a property of the op-amp. It's normally absurdly large; for example, the gain of the LM4250 is specified to be at least 25 000. This gain is also, generally, not well-defined. Manufacturers tend to guarantee a *minimum* (25 000 in the case of the LM4250), and *maybe* a maximum, but the range will be too large for you to rely on even an approximate value.

You might then wonder: What use is this device, that multiplies a difference between two inputs by a number to big for me to imagine, and that I can't even rely on being a specific value? Great

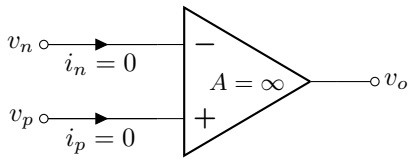


Figure 8.3: Ideal op-amp

question. The key is to put the device in a circuit that uses *feedback* to set the gain to a reasonable value. To better understand what feedback is, and how we use it in op-amp circuits, we will create an idealized model of the op-amp—you might have noticed this theme in this course—known as the *ideal op-amp*.

8.1.1 The Ideal Op-Amp and Negative Feedback

The *ideal* op-amp has *infinite gain* and *zero input current* (Figure 8.3):

$$A = \infty \tag{8.2}$$
$$i_p = 0 \tag{8.3}$$
$$i_n = 0 \tag{8.4}$$

This probably seems even weirder. Before we were talking about A being a really large number, now we're saying the gain is *infinite*?! Doesn't that mean that $v_o = \infty$, always? Perhaps surprisingly, no, depending on how you use the op-amp. Specifically, the output ends up being finite if you use the op-amp in a configuration that has *negative feedback*.

Feedback systems

Feedback is when you connect the output of a circuit back to its input, so that the output "feeds back" into the circuit. Many other systems also have feedback—if you're sensing a quantity in order to control it, you have some form of a feedback system.

For example, many air conditioning systems use temperature sensors to determine how much power they should exert in heating or cooling the room: if the sensed temperature is too high, it applies cooling, and if it's too low, it applies heating. Ovens and refrigerators both use similar ideas to control their temperature (except that an oven's idea of "cooling" and a fridge's idea of "heating" is to do nothing).

The useless box you built in lab 2 uses *feedback*, of a sort: it detects the state of the toggle switch, in order to activate a finger that changes the state of the toggle switch. You might not realize it, but in every day situations, *you* are a feedback system. For example, when you're driving a car, you're (hopefully) watching your speedometer, and using what you read to inform how hard you press the accelerator.

Negative feedback

When we design feedback so that an *increase* in the output is fed back to the input to cause a *decrease* in the output, and a *decrease* in the output causes an *increase* in the output, we call it *negative feedback*. Negative feedback is like self-correction—when you detect yourself going too high, you pull yourself back down, and vice versa. The temperature control and speed control systems we just described are both negative feedback systems. When you're driving, if the speedometer shows a speed higher than what you want, you'll lighten up on the accelerator; if it's slower, you'll press harder.

Voltage follower

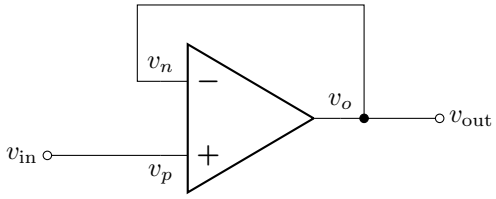


Figure 8.4: Voltage follower

Now, back to op-amp circuits. Consider the circuit in Figure 8.4, and think of the op-amp's gain A as a very large (but finite) number. In this circuit, $v_{\text{out}} = v_o = v_n$, because we connected the output directly to the inverting input. Then, if the output voltage v_{out} increases, then $v_p - v_n$ decreases, which in turn means $A(v_p - v_n)$ decreases. So, the fact that the output voltage goes *up*, causes the output voltage to want to go *down*. In other words, the circuit is in negative feedback.

You might wonder: But if A is *really* large, then won't v_o oscillate wildly between very positive and very negative, as v_n goes above and below v_p? As it happens, v_o can't change instantaneously—it changes very quickly, but it still has to be continuous. That is, a new v_o doesn't magically appear at the output; rather, v_o moves *towards* its new value.

The next question is, then, when does it stop? A mathematically robust approach would take a finite gain A, find the point where v_o would be consistent with itself (so feedback has reached an equilibrium), and take the limit as $A \to \infty$; we explore this approach in Section 8.1.4.

Why $v_p = v_n$

A more hand-waving (but still true) argument is as follows: If the gain A is infinite (extremely large), then the only way for $v_o = A(v_p - v_n)$ to be finite (not extremely large) is if $v_p - v_n = 0$ (extremely small). At the limit, then, if the op-amp is in a circuit with negative feedback,

$$v_p = v_n. \tag{8.5}$$

Golden rules

Equations (8.2) through (8.5) are collectively known as the *golden rules of ideal op-amps in negative feedback*. For convenience, these are summarized in the box below.

Golden rules of ideal op-amps in negative feedback

$$A = \infty \qquad i_p = 0$$
$$v_p = v_n \qquad i_n = 0$$

In the case of the circuit in Figure 8.4, we'll have $v_{\text{in}} = v_p = v_n = v_{\text{out}}$. So the output is just equal to the input. For this reason, that circuit is often called a *voltage follower*, or a *voltage buffer*.

Problem 8.1 How does the voltage follower circuit in Figure 8.4 differ from just having a short circuit from v_{in} to v_{out}? In what situations might you want to use it?

8.1.2 Output Saturation

In principle, an ideal op-amp would output whatever voltage is asked of it; in reality, it can only output voltages within the range of its power supply. That is, if the positive power supply terminal is connected to V_{DD}, and the negative power supply terminal is connected to $-V_{DD}$, then only

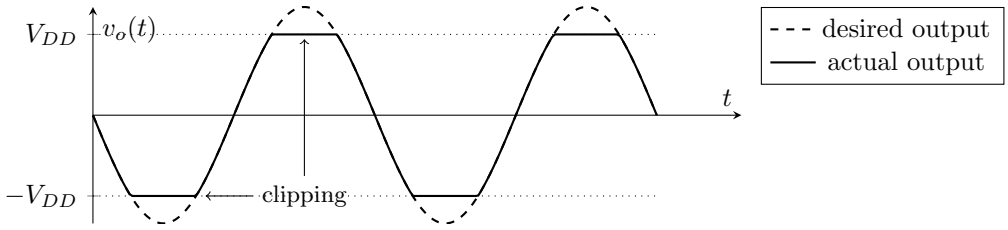


Figure 8.5: Clipped signal

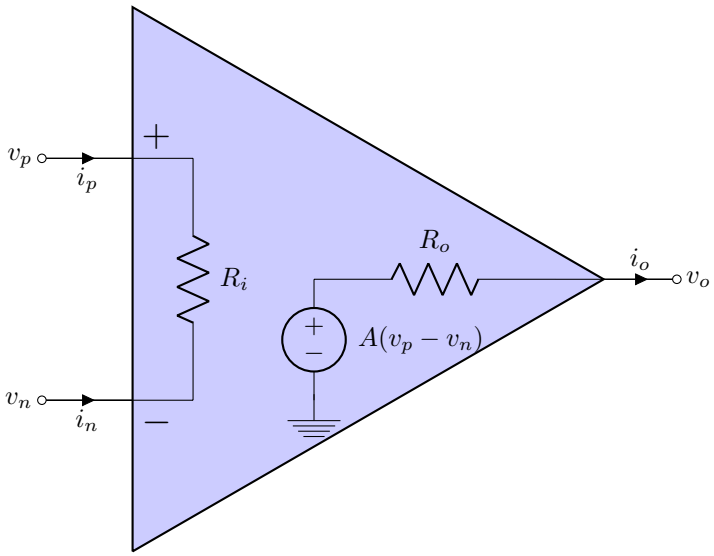


Figure 8.6: Equivalent circuit model of an operational amplifier

outputs between $-V_{DD}$ and V_{DD} are possible. When the desired output voltage would exceed the possible range, the op-amp is said to *saturate*, and the op-amp just outputs its maximum or minimum possible voltage instead.

We often call the supply voltages the *rails*. When op-amp output saturation causes the signal to be cut off close to the rails, limit by its inability to output outside the supply voltage range, we say that the signal is *clipped*. An illustration of this is in Figure 8.5.

8.1.3 Output Drive, and Input Current

The ideal op-amp we described has $i_p = 0$ and $i_n = 0$, and the output voltage $v_o = A(v_p - v_n)$ exactly. A more realistic model of an op-amp would acknowledge a small current at the inputs, and internal resistance at the output. An equivalent circuit model of an op-amp showing one such model is shown in Figure 8.6.

Ideal values for internal resistance

If the op-amp were ideal, then we would need $i_p = 0$ and $i_n = 0$, which implies that $R_i = \infty$. Also, for the output v_o to be exactly $A(v_p - v_n)$, we would need $R_o = 0$. For this reason, you'll often see other texts say that $R_i = \infty$ and $R_o = 0$ are included in the *golden rules of ideal op-amps*

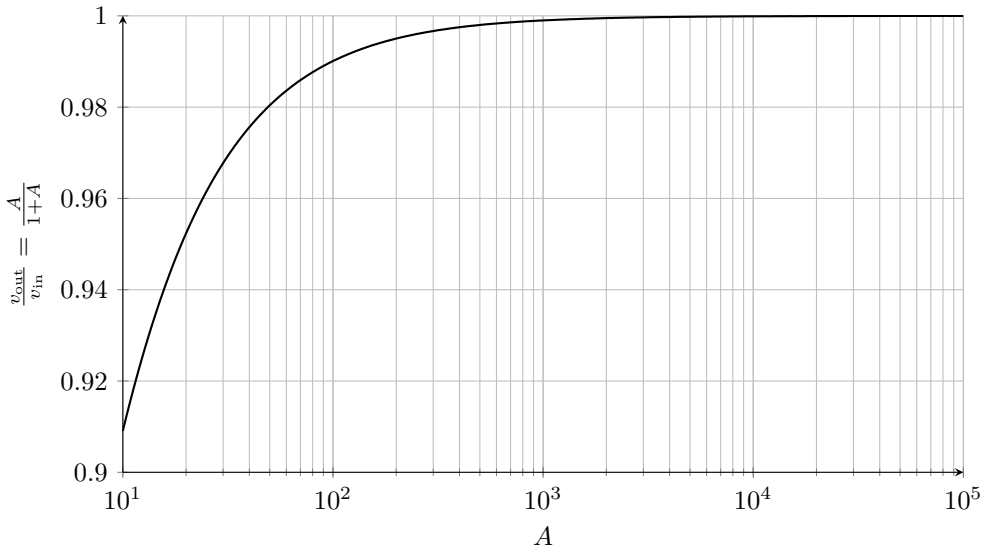


Figure 8.7: Gain of voltage follower in circuit of Figure 8.4 vs op-amp gain A

that we outlined in section 8.1.1. In most applications, R_i is sufficiently large and R_o sufficiently small that the ideal op-amp is a good approximation.

Output is a voltage source

The model in Figure 8.6 also demonstrates a good way to think about the output: as a voltage source that *depends on* v_p and v_n. As v_p or v_n change, the voltage of this source changes too. In future classes, we'll make formalize this concept as a "dependent voltage source". For now, the salient point is that, like a voltage source, the output of an op-amp will provide whatever current is necessary to maintain the voltage $A(v_p - v_n)$ relative to ground. In an ideal op-amp with $A = \infty$, this is whatever voltage is necessary to make $v_p = v_n$.

8.1.4 Circuits with Finite-Gain Op-Amps

In circuits with ideal op-amps, the approximation $A = \infty$ allows us to conclude that, in a circuit with negative feedback, $v_p = v_n$. What happens in a circuit with finite gain? Consider the voltage follower in Figure 8.4 again, but this time, let A be finite.

Follower with finite gain

Recall that $v_o = A(v_p - v_n)$, and in this circuit, $v_n = v_o = v_{\text{out}}$ and $v_p = v_{\text{in}}$. Then, when v_o is consistent with itself, we'll have

$$\begin{aligned} v_o &= A(v_p - v_n) \qquad (8.6)\\ v_{\text{out}} &= A(v_{\text{in}} - v_{\text{out}}) \\ v_{\text{out}} &= Av_{\text{in}} - Av_{\text{out}} \\ v_{\text{out}}(1 + A) &= Av_{\text{in}} \\ v_{\text{out}} &= \frac{A}{1+A} v_{\text{in}}. \qquad (8.7) \end{aligned}$$

For example, if you have a particularly low-gain op-amp of just $A = 100$, with this circuit you'll

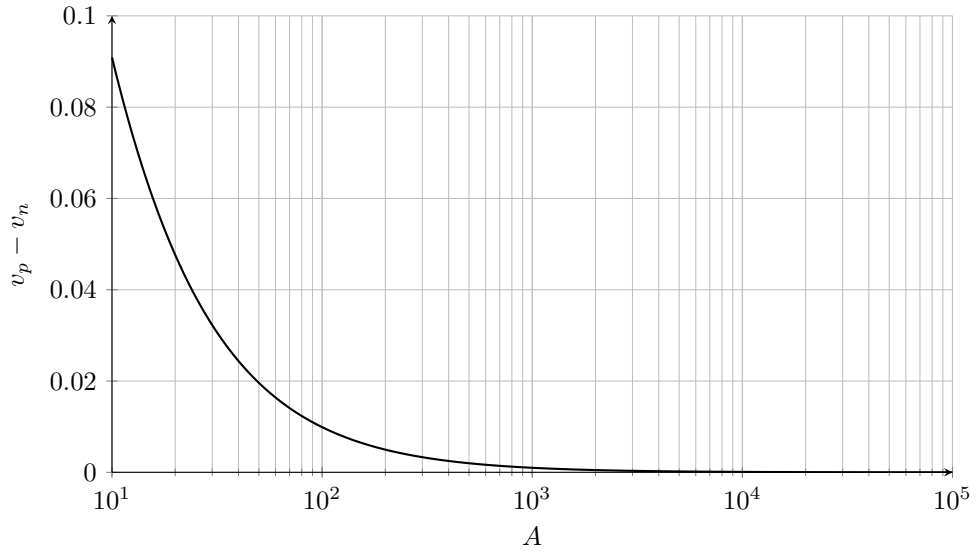


Figure 8.8: Op-amp input difference $v_p - v_n$ in circuit of Figure 8.4 vs op-amp gain A, with $v_{\text{in}} = 1\,\text{V}$

get $v_{\text{out}} = \frac{100}{101} v_{\text{in}}$. A graph of the expression $\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{A}{1+A}$, with A on a logarithmic scale, is shown in Figure 8.7.

Behavior as $A \to \infty$

Notice that as the A increases, the gain of the circuit $\frac{v_{\text{out}}}{v_{\text{in}}}$ asymptotically approaches its ideal value, 1. In the limit, as $A \to \infty$, $v_p - v_n \to 0$ and $v_{\text{out}} \to v_{\text{in}}$. This gives some idea of why it's valid to use $v_p = v_n$ when working with ideal op-amps in negative feedback.

In fact, we can make this idea more precise. It follows from (8.6) and (8.7) that

$$v_p - v_n = \frac{1}{1+A} v_{\text{in}}. \tag{8.8}$$

A graph of this against A is shown in Figure 8.8, and it can be seen that it approaches zero as $A \to \infty$.

8.2 Basic Op-Amp Circuits

The voltage follower of Figure 8.4 is a common use of op-amps. In this section, we outline a two other basic, and extremely common, uses of op-amps.

8.2.1 Non-inverting Amplifier

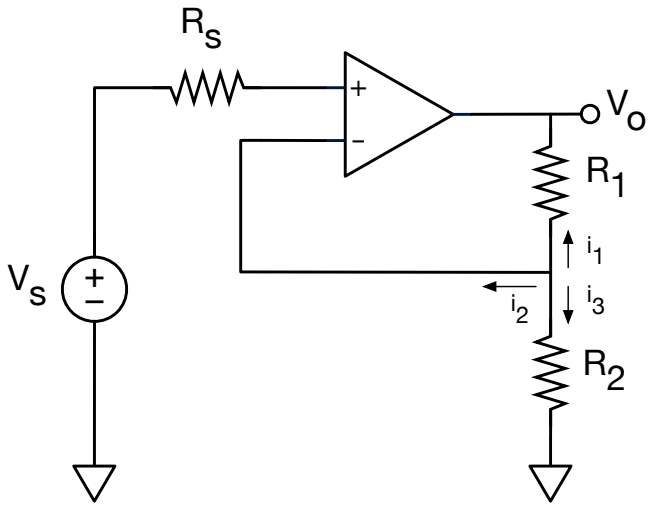


Figure 8.9: Noninverting amplifier circuit using the ideal op-amp model.

A noninverting amplifier can be used to amplify an input signal without inverting its sign. Figure 8.9 depicts the noninverting amplifier using an ideal op-amp model. The output drives a resistor divider, and the output of this resistor divider is connected to the negative input of the op-amp. Using this the ideal op-amp model, the analysis is relatively straightforward.

The new circuit contains a source resistor that connects from the voltage source v_s to the positive input terminal of the op-amp. However, since $i_p = 0$, there is no voltage drop across it and the voltage at $v_p = v_s$. Next, we will write a KCL equation at v_n:

$$i_1 + i_2 + i_3 = 0 \tag{8.9}$$

We know $i_2 = i_n = 0$ so we don't need to worry about that. We also know that the op-amp will find the output voltage that makes vn = vp, so to find this output voltage, we can set $v_n = v_p$, which means that $v_n = v_s$ (these are the "golden rules"). Now we have everything we need to solve the problem. First we obtain the following equation by substituting the expressions for i_1, i_2, and i_3.

$$\begin{aligned} \frac{v_s - v_o}{R_1} + \frac{v_s}{R_2} &= 0 \\ \frac{v_o}{R_1} &= \frac{v_s}{R_1} + \frac{v_s}{R_2} \\ \frac{v_o}{R_1} &= v_s \left(\frac{R_1 + R_2}{R_1 R_2} \right) \end{aligned}$$

Therefore, we are left with the following gain expression:

$$\frac{v_o}{v_s} = \frac{R_1 + R_2}{R_2} \tag{8.10}$$

8.2.2 Inverting Amplifier

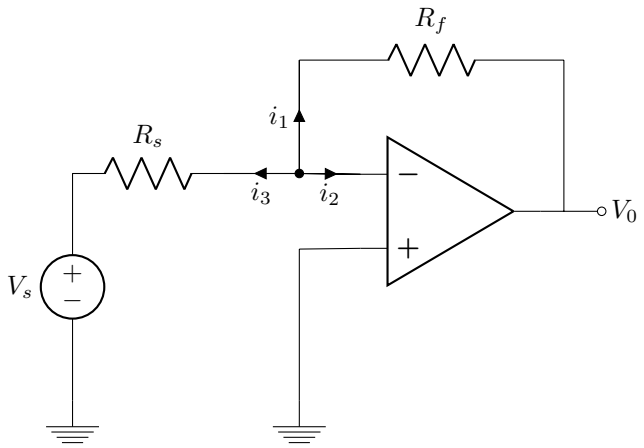


Figure 8.10: Inverting amplifier using the ideal op-amp model

As with the noninverting amplifier, the analysis of the inverting amplifier is significantly facilitated using the ideal op-amp model. In the inverting amplifier, notice that the input source is connected to terminal v_n through a resistor R_s and the terminal v_p is connected to ground. The resistor R_f allows the output to be applied continuously to the input terminal or "feed back" into the input terminal via the resistor R_f.

We can start our analysis of this circuit by writing a KCL equation at the negative terminal v_n.

$$i_1 + i_2 + i_3 = 0 \tag{8.11}$$

By inserting the definitions of i_1, i_2, and i_3, this equation can be rewritten as:

$$\frac{v_n - v_s}{R_s} + \frac{v_n - v_o}{R_f} + i_n = 0 \tag{8.12}$$

Since $i_n = 0$ and $v_n = v_p = 0$ (using the "golden rules"), the above equation can be simplified to:

$$\frac{0 - v_s}{R_s} + \frac{0 - v_o}{R_f} + 0 = 0$$

$$\frac{-v_s}{R_s} - \frac{v_o}{R_f} = 0$$

Therefore, we have

$$\frac{v_o}{v_s} = -\left(\frac{R_f}{R_s}\right) \tag{8.13}$$

Since the gain is negative, this circuit is referred to as an inverting amplifier.

8.3 Other Useful Amplifier Circuits

Operational amplifiers have various applications as building blocks in analog signal processing circuits. In this section, we will discuss four operations that can be performed using op-amps: addition, subtraction, low-pass filtering, and high-pass filtering.

8.3.1 Summing Amplifier

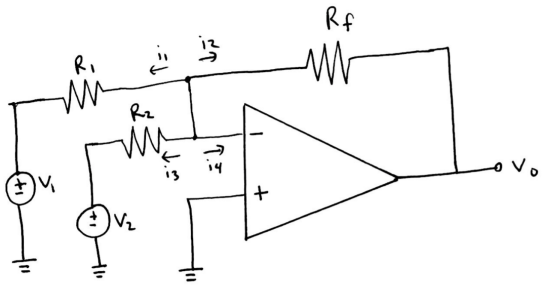


Figure 8.11: Summing amplifier

If you think back at the inverting amplifier circuit in Figure 8.10, you might notice that the source resistor, R_s converts the input voltage into a current (since v_n is kept at zero voltages, and then this current is converted back to a voltage using R_f. If we connected multiple inputs to v_n, as shown in Figure 8.11, the output voltage will depend on the sum of the input currents. Since each current is equal to the input voltage divided by its source resistor, the output will be a weighted sum of the input voltages. The inputs are weighted by $1/R_s$. The precise expression can be derived by first writing a KCL equation at the negative input terminal:

$$i_1 + i_2 + i_3 + i_4 = 0 \tag{8.14}$$

By substituting the definitions of i_1, i_2, i_3, and i_4, we have the following equation:

$$\frac{v_n - v_1}{R_1} + \frac{v_n - v_o}{R_f} + \frac{v_n - v_2}{R_2} + i_n = 0 \tag{8.15}$$

By applying the "golden rules," we note that $i_n = 0$ and $v_n = v_p = 0$, and we arrive at the following equation:

$$-\frac{v_1}{R_1} - \frac{v_o}{R_f} - \frac{v_2}{R_2} = 0$$

$$\frac{v_o}{R_f} = -\frac{v_1}{R_1} - \frac{v_2}{R_2}$$

Therefore, we can write v_o as:

$$v_o = -\left(\frac{R_f}{R_1}\right) v_1 - \left(\frac{R_f}{R_2}\right) v_2 \tag{8.16}$$

Since the gain for each of the input sources is negative, this circuit is occasionally referred to as a "scaled inverting adder." In the special case where $R_1 = R_2 = R_f$, we have the inverted sum of the input voltages:

$$v_o = -(v_1 + v_2) \tag{8.17}$$

Furthermore, we can easily see that additional sources can be added to the negative terminal v_n. In the case where we have n sources with source resistances ranging from R_1 to R_n, the output voltage becomes:

$$v_o = -\left(\frac{R_f}{R_1}\right) v_1 - \left(\frac{R_f}{R_2}\right) v_2 - ... - \left(\frac{R_f}{R_n}\right) v_n \tag{8.18}$$

8.3.2 Difference Amplifier

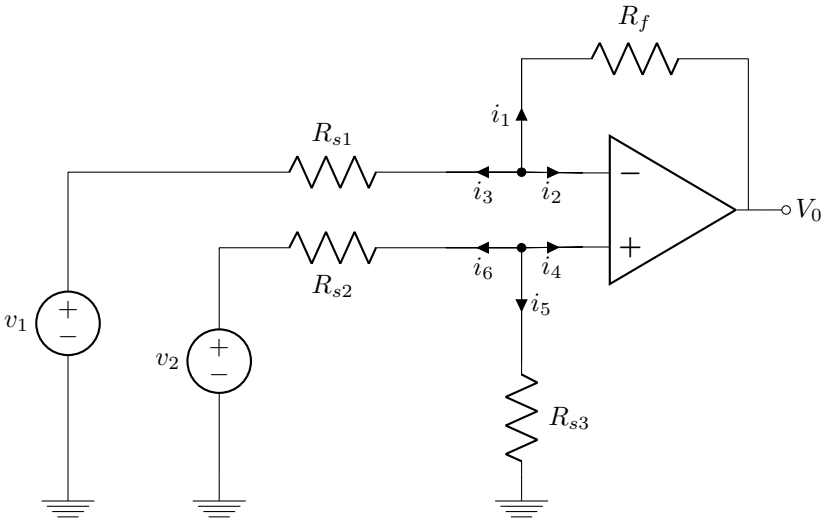


Figure 8.12: Difference amplifier

In the summing amplifier, the output voltage depended on two (or more) input voltages with the same sign. In the difference amplifier, the input sources v_1 and v_2 will have opposite polarities at the output. Therefore, this circuit can be used to perform the subtraction of two signals. The circuit schematic of a difference amplifier can be seen in Figure 8.12.

In order to derive a relationship for the output voltage, we first write a KCL equation at the negative terminal.

$$i_1 + i_2 + i_3 = 0 \tag{8.19}$$

By substituting the definitions of i_1, i_2, and i_3, we obtain:

$$\frac{v_n - v_o}{R_f} + i_n + \frac{v_n - v_1}{R_{s1}} = 0 \tag{8.20}$$

Next, we write a KCL equation at the positive terminal:

$$i_4 + i_5 + i_6 = 0 \tag{8.21}$$

By substituting the definitions of i_4, i_5, and i_6, we obtain:

$$i_p + \frac{v_p}{R_{s3}} + \frac{v_p - v_2}{R_{s2}} = 0 \tag{8.22}$$

By noticing that $i_p = i_n = 0$ and $v_p = v_n$, we write the following two equations for this circuit:

$$\frac{v_n - v_o}{R_f} + \frac{v_n - v_1}{R_{s1}} = 0 \Longrightarrow R_{s1} \cdot (v_n - v_o) + R_f \cdot (v_n - v_1) = 0 \tag{8.23}$$

$$\frac{v_n}{R_{s3}} + \frac{v_n - v_2}{R_{s2}} = 0 \Longrightarrow R_{s2} \cdot v_n + R_{s3} \cdot (v_n - v_2) = 0 \tag{8.24}$$

Solving the second equation gives v_n as a function of v_2:

$$v_n = \frac{R_{s3}}{R_{s2} + R_{s3}} v_2 \tag{8.25}$$

This makes sense, since v_n is driven by a voltage divider formed by R_{s2} and R_{s3}. This can then be used to find v_o

$$v_o = \left(\frac{R_f + R_{s1}}{R_{s1}}\right)\left(\frac{R_{s3}}{R_{s3} + R_{s2}}\right) v_2 - \left(\frac{R_f}{R_{s1}}\right) v_1 \tag{8.26}$$

In order for the difference amplifier to subtract the two voltages with equal gain, the resistors must be related via the following equation:

$$R_f R_{s2} = R_{s3} R_{s1} \tag{8.27}$$

If this equation is satisfied, then the output voltage can be expressed as:

$$v_o = \left(\frac{R_f}{R_{s1}}\right)(v_2 - v_1) \tag{8.28}$$

8.3.3 Active High-pass Filter

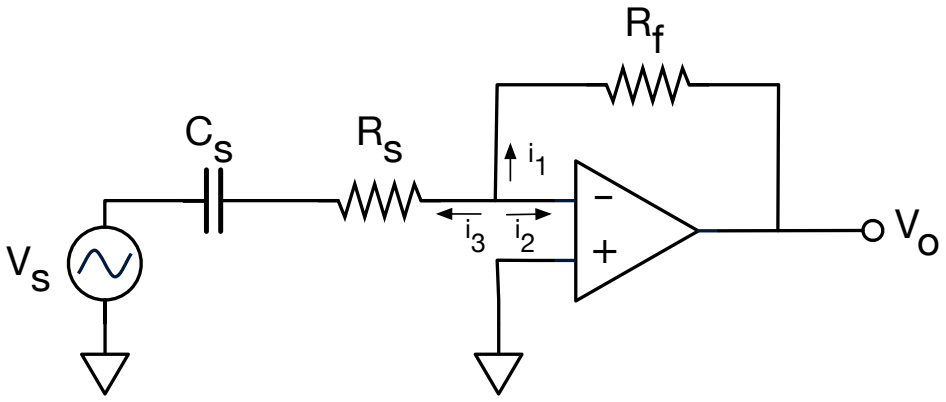


Figure 8.13: Active High-pass Filter

In many applications, we want to amplify an input signal over a specific range of frequencies while attenuating the signal over another range of frequencies. A filter that uses active components, such as transistors, op-amps, voltage sources, current sources, etc. is referred to as an "active filter." One advantage of an active filter as opposed to a passive filter (a filter that only consists of passive elements: resistors, inductors, and capacitors) is that an active filter can amplify the signal while simultaneously filtering out unwanted frequency components. In this section, we will analyze an active high-pass filter. This filter amplifiers the high frequency components of a signal (that is, the frequencies above the cut-off frequency) and attenuates low frequencies of the signal (that is, the frequencies that are below the cut-off frequency). The active high pass filter can be seen in Figure 8.13. Notice that the input signal is an AC voltage source.

In order to simplify the analysis of this circuit, we first transform the components C_s and R_s into an equivalent impedance element Z_s. Then, we transform R_f into an equivalent impedance element Z_f. This transformation can be seen in Figure 8.14. Since C_s and R_s are in series, we have:

$$Z_s = \frac{1}{j \cdot 2\pi f C} + R_s \tag{8.29}$$

Since R_f is a resistor:

$$Z_f = R_f \tag{8.30}$$

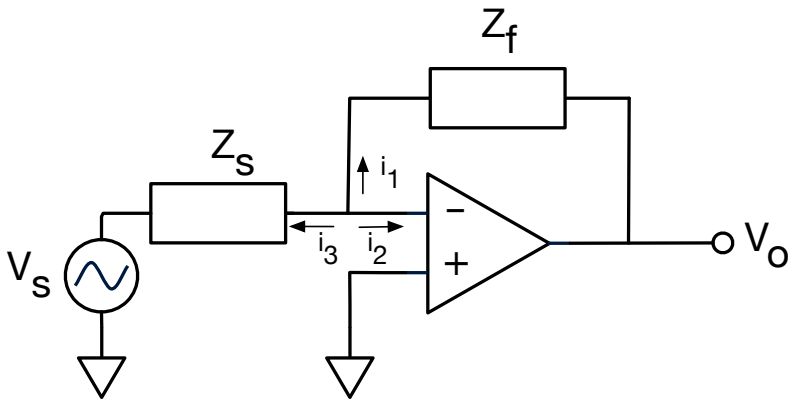


Figure 8.14: Inverting amplifier with complex impedances

The amplifier in Figure 8.14 looks almost identical to the inverting amplifier that we analyzed earlier. However, instead of resistors we have impedances. We analyze this circuit in the same manner as the inverting amplifier by first writing a KCL equation at the negative terminal of the op-amp:

$$i_1 + i_2 + i_3 = 0 \tag{8.31}$$

Next, we substitute in the definitions of i_1, i_2, and i_3:

$$\frac{v_n - v_o}{Z_f} + i_n + \frac{v_n - v_s}{Z_s} = 0 \tag{8.32}$$

We now apply the the conditions that $i_n = 0$ and $v_p = v_n = 0$ to obtain:

$$\frac{-v_o}{Z_f} + \frac{-v_s}{Z_s} = 0 \tag{8.33}$$

Next, we solve for the gain:

$$\frac{-v_o}{Z_f} + \frac{-v_s}{Z_s} = 0$$

$$\frac{-v_o}{Z_f} = \frac{v_s}{Z_s}$$

$$\frac{v_o}{v_s} = -\frac{Z_f}{Z_s} \tag{8.34}$$

Finally, we substitute in the definitions of Z_f and Z_s and simplify:

$$\frac{v_o}{v_s} = -\frac{R_f}{\frac{1}{j2\pi f C_s} + R_s} \tag{8.35}$$

$$\frac{v_o}{v_s} = -\frac{j \cdot 2\pi f R_f C_s}{1 + j \cdot 2\pi f R_s C_s} \tag{8.36}$$

From equation 8.36, we notice that the gain of the amplifier at high frequencies is $-\frac{R_f}{R_s}$. This result agrees with the gain for an inverting amplifier. This finding is as we would expect, since at high frequencies the impedance of a capacitor decreases, thus the behavior of the circuit in Figure 8.13 would approach the behavior of the circuit shown in Figure 8.9. The Bode plot for the active filter is shown in Figure 8.15. The following parameters were used in the simulation: $R_f = 10\,\text{k}\Omega$, $R_s = 1\,\text{k}\Omega$, and $C_s = 1\,\mu\text{F}$.

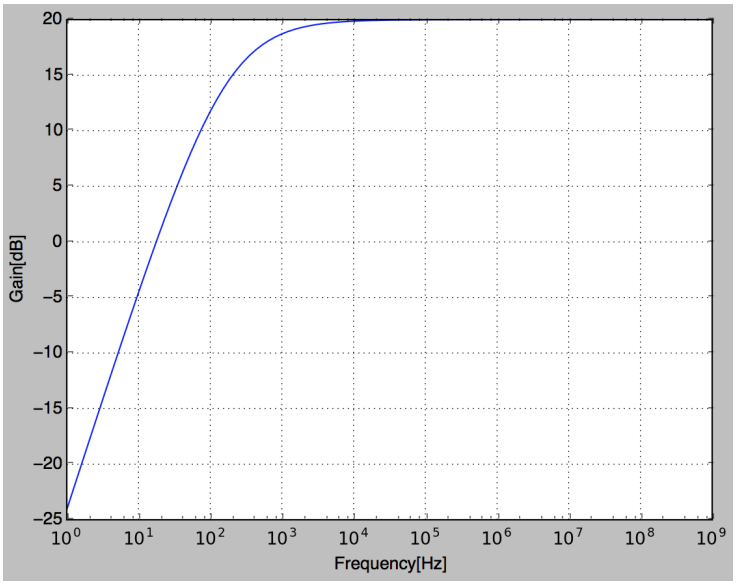


Figure 8.15: Bode Plot for Active High Pass Filter

Notice that the magnitude of the gain of the circuit at high frequencies is equal to: $\frac{v_o}{v_s} = \frac{R_f}{R_s} = 10$. This corresponds to 20dB. Additionally, notice that the cutoff frequency is:

$$f_c = \frac{1}{2\pi R_s C_s} \tag{8.37}$$

Therefore, in this example, $f_c \approx 159$Hz

8.3.4 Active Low-pass Filter

We can also design an active version of the low-pass filter. This filter amplifiers the low frequency components of the signal (that is, the frequencies below the cut-off frequency) and attenuates the high frequency components of the signal (that is, the frequencies that are above the cut-off frequency). The active low-pass filter can be sen in Figure 8.16. Notice that the input signal is an AC voltage source.

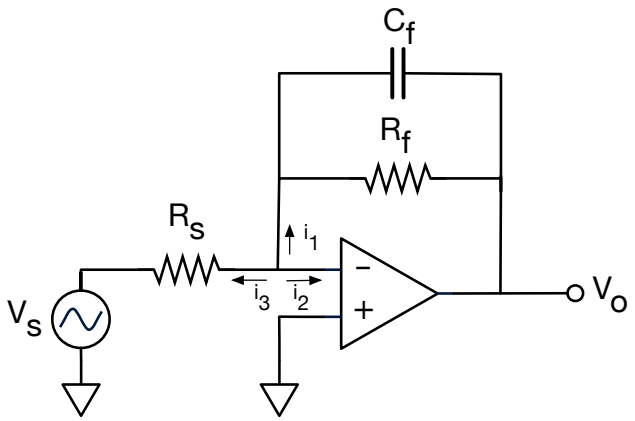


Figure 8.16: Active Low-pass Filter

In order to simply the analysis of this circuit, we first transform the components C_f and R_f into an equivalent impedance element Z_f. Then, we transform R_s into an equivalent impedance element Z_s. This transformation can be seen in Figure 8.17. Since C_f and R_f are in parallel, we have[1]:

$$Z_f = \frac{1}{j2\pi f C_f} \parallel R_f \tag{8.38}$$

$$Z_f = \frac{R_f}{1 + j2\pi f R_f C_f} \tag{8.39}$$

Since R_s is a resistor:

$$Z_s = R_s \tag{8.40}$$

[1]Notice that we are using the "no phase approximation"

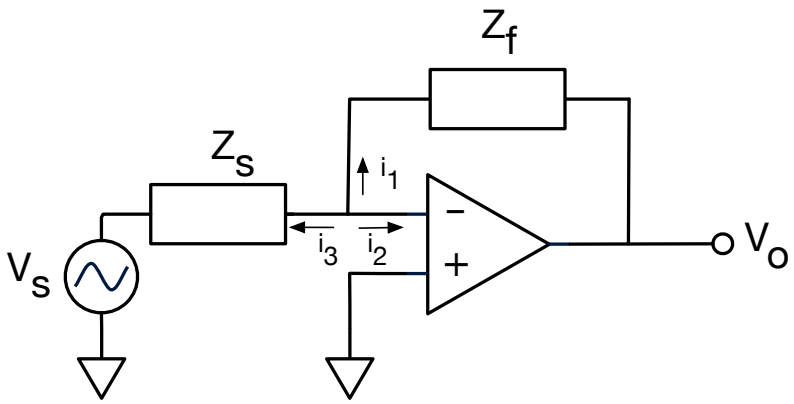


Figure 8.17: Inverting amplifier with complex impedances

The amplifier in Figure 8.17 looks almost identical to the inverting amplifier that we analyzed earlier. However, instead of resistors we have impedances. We analyze this circuit in the same manner as the inverting amplifier by first writing a KCL equation at the negative terminal of the op-amp:

$$i_1 + i_2 + i_3 = 0 \tag{8.41}$$

Next, we substitute in the definitions of i_1, i_2, and i_3:

$$\frac{v_n - v_o}{Z_f} + i_n + \frac{v_n - v_s}{Z_s} = 0 \tag{8.42}$$

We now apply the the conditions that $i_n = 0$ and $v_p = v_n = 0$ to obtain:

$$\frac{-v_o}{Z_f} + \frac{-v_s}{Z_s} = 0 \tag{8.43}$$

Next, we solve for the gain:

$$\frac{-v_o}{Z_f} + \frac{-v_s}{Z_s} = 0$$

$$\frac{-v_o}{Z_f} = \frac{v_s}{Z_s}$$

$$\frac{v_o}{v_s} = -\left(\frac{Z_f}{Z_s}\right) \tag{8.44}$$

Finally, we substitute in the definitions of Z_f and Z_s and simplify:

$$\frac{v_o}{v_s} = -\left(\frac{\frac{R_f}{1+j2\pi f R_f C_f}}{R_s}\right) \tag{8.45}$$

$$\frac{v_o}{v_s} = -\left(\frac{R_f}{R_s}\right)\left(\frac{1}{1+j2\pi f R_f C_f}\right) \tag{8.46}$$

In one example of an active low-pass filter, let $R_f = 10\,\text{k}\Omega$, $R_s = 1\,\text{k}\Omega$, and $C_f = 1\mu F$. The magnitude of the gain of the active filter at low frequencies is equal to:

$$\frac{v_o}{v_s} = \frac{R_f}{R_s} = 10 \tag{8.47}$$

A voltage gain of 10 corresponds to a gain of 20dB. The cutoff frequency is:

$$f_c = \frac{1}{2\pi R_f C_f} = 15.9\text{Hz} \tag{8.48}$$

The results of this derivation can be seen in the following Bode plot for the transfer function of the active low-pass filter:

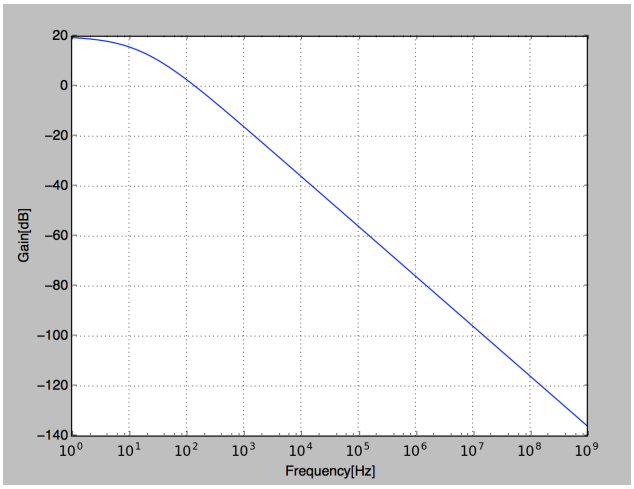


Figure 8.18: Bode plot for the low-pass active filter

8.4 Additional Applications

8.4.1 Voltage Follower

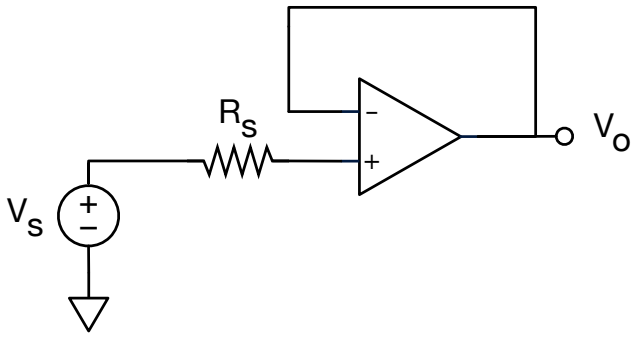


Figure 8.19: Voltage follower

The voltage follower as shown in Figure 8.19 is a ubiquitous circuit that is used to isolate the input signal from variations in the output circuit. From a simple application of the op-amp "golden rules", we find that:

$$v_s = v_p \text{ (Since } i_p = 0) \tag{8.49}$$

We also find that:

$$v_p = v_n = v_o \tag{8.50}$$

Therefore:

$$v_s = v_o \tag{8.51}$$

From the equation above it is clear why this circuit is referred to as a voltage follower since the output voltage "follows" the input voltage.

However, what is the point of this circuit if it doesn't appear to do any kind of operation on the input signal? The utility the voltage follower can be seen by examining Figure 8.20 and Figure 8.21. In Figure 8.20, the ouput voltage is determined by the voltage divider equation:

$$v_o = v_s \left(\frac{R_L}{R_s + R_L} \right) \tag{8.52}$$

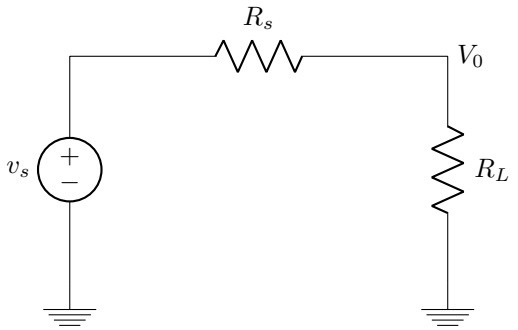


Figure 8.20: Voltage divider

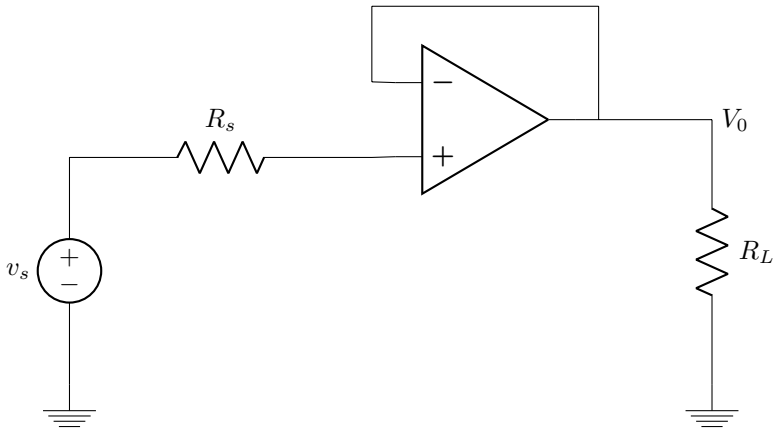


Figure 8.21: Voltage follower as a buffer

Notice that the input signal V_s is attenuated by the voltage divider. Suppose, however, that this attenuation is undesirable and that we want the entire input signal V_s to appear across the load. By inserting a voltage follower in between R_s and R_L, we can accomplish just that! Since the current $i_p = 0$, the voltage $v_p = v_s$. Therefore, we have

$$v_s = v_p = v_n = v_o \tag{8.53}$$

8.4.2 Instrumentation Amplifier

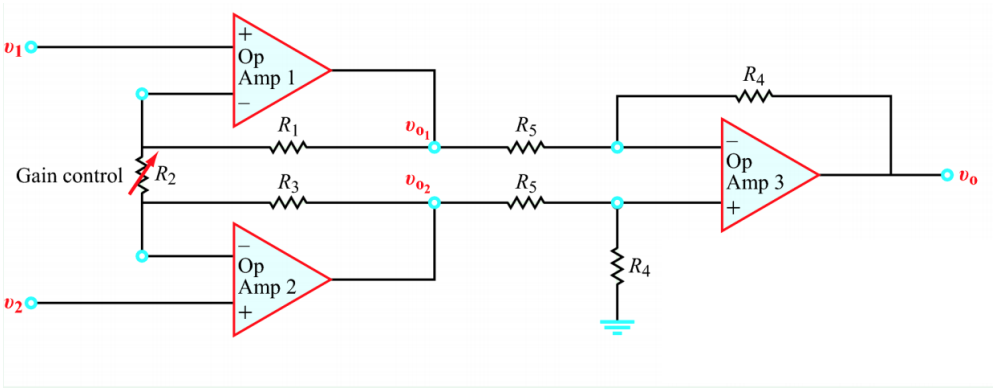


Figure 8.22: Instrumentation amplifier

An instrumentation amplifier as shown in Figure 8.22 is used to amplify a small difference between two signals. This type of circuit is often used in sensors to detect deviations from a nominal value. If one can relate a physical quantity such as temperature, pressure, humidity, etc. to a voltage, then by using an instrumentation amplifier a sensor can be designed to detect when, for example, room temperature/humidity/pressure/etc. deviates from an acceptable value. In Engr40M, we will be using an instrumentation amplifier as part of a larger circuit that can monitor your heart rate!

To express v_o in terms of v_1 and v_2, we first note that the the third amplifier is a difference amplifier with inputs v_{o1} and v_{o2}. Therefore, using the result we derived earlier, we have the following equation:

$$v_o = -\left(\frac{R_4}{R_5}\right) v_{o1} + \left(\frac{R_4 + R_5}{R_5}\right)\left(\frac{R_4}{R_4 + R_5}\right) v_{o2} \tag{8.54}$$

This can be simplified to:

$$v_o = \left(\frac{R_4}{R_5}\right)(v_{o2} - v_{o1}) \tag{8.55}$$

Next, we determine the intermediate output voltages v_{o1} and v_{o2}. We first write a KCL equation at v_{n1}, the negative input of the first amplifier and apply the op-amp "golden rules:"

$$\frac{v_1 - v_2}{R_2} + \frac{v_1 - v_{o1}}{R_1} = 0 \tag{8.56}$$

This can be simplified to:

$$v_{o1} = v_1 + \frac{R_1}{R_2}(v_1 - v_2) \tag{8.57}$$

Next, we write a KCL equation at v_{n2}, the negative input of the second amplifier and apply the op-amp "golden rules:"

$$\frac{v_2 - v_1}{R_2} + \frac{v_2 - v_{o2}}{R_3} = 0 \tag{8.58}$$

This can be simplified to:

$$v_{o2} = v_2 + \frac{R_3}{R_2}(v_2 - v_1) \tag{8.59}$$

Next, we subtract equation (8.57) from equation (8.59) to obtain:

$$v_{o2} - v_{o1} = v_2 - v_1 + \frac{R_3}{R_2}(v_2 - v_1) + \frac{R_1}{R_2}(v_2 - v_1) \tag{8.60}$$

This can be simplified to:

$$v_{o2} - v_{o1} = \left(\frac{R_1 + R_2 + R_3}{R_2}\right)(v_2 - v_1) \tag{8.61}$$

Finally, we can obtain the overall expression for v_o in terms of v_1 and v_2 by substituting equation (8.61) into equation (8.55), and we obtain the following:

$$v_o = \left(\frac{R_4}{R_5}\right)\left(\frac{R_1 + R_2 + R_3}{R_2}\right)(v_2 - v_1) \tag{8.62}$$

In the case where all the resistors are equal except for R_2, the expression for the output voltage becomes:

$$v_o = \left(1 + \frac{2R}{R_2}\right)(v_2 - v_1) \tag{8.63}$$

In this case, the gain of the instrumentation amplifier is set by R_2. This configuration can be very useful if one desires to have a variable gain since R_2 can be be implemented as a potentiometer.

8.5 Summary

The Golden Rules for Ideal Op-Amps

- $A = \infty$
- $R_i = \infty$
- $R_o = 0$
- $i_p = i_n = 0$
- $v_n = v_p$ (for negative feedback configurations)

The output voltage of an op-amp is limited by the supply voltage for the op-amp:

$$|v_o| \leq V_{dd} \tag{8.64}$$

8.6 Analysis of non-ideal Op Amp (bonus material)

The previous sections analyzed an ideal op-amp behavior. Analyzing the behavior of a non-ideal op-amp is only a little more difficult and is done in the next sections.

8.6.1 Noninverting Amplifier

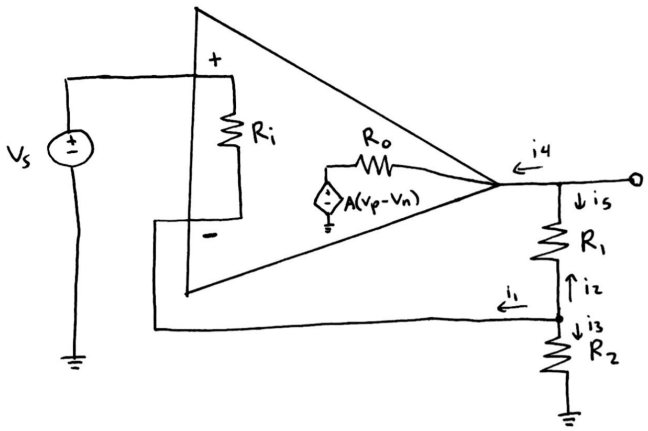


Figure 8.23: A noninverting amplifier using the equivalent circuit model of an op-amp

A noninverting amplifier can be used to amplify an input signal without inverting its sign. In order to derive an expression for the gain of this amplifier, we write a KCL equation at the negative input of the amplifier and a KCL equation at the output of the amplifier:

KCL @ v_n:

$$i_1 + i_2 + i_3 = 0 \tag{8.65}$$

Next, we substitute in the expressions for i_1, i_2, and i_3, and we note that $v_p = v_s$ for this circuit. After making these substitutions, we have the following equation:

$$\frac{v_n - v_s}{R_i} + \frac{v_n - v_o}{R_1} + \frac{v_n}{R_2} = 0 \tag{8.66}$$

KCL @ v_o:

$$i_4 + i_5 = 0 \tag{8.67}$$

Next, we substitute in the expressions for i_4 and i_5, and we note that $v_p = v_s$ for this circuit. After making these substitutions, we have the following equation:

$$\frac{v_o - A(v_s - v_n)}{R_o} + \frac{v_o - v_n}{R_1} = 0 \tag{8.68}$$

By combining equations (8.66) and (8.68) and through a painful amount of algebra, we obtain the following expression for the gain $\left(\frac{v_o}{v_s}\right)$ of the circuit:

$$\frac{v_o}{v_s} = \frac{AR_i\left(R_1 + R_2\right) + R_2R_0}{AR_2R_i + R_o\left(R_2 + R_i\right) + R_1R_2 + R_i\left(R_1 + R_2\right)} \tag{8.69}$$

Suppose that for one particular application, $R_1 = 90\,\text{k}\Omega$, $R_2 = 10\,\Omega$, $A = 10^6$, $R_i = 10\,\text{M}\Omega$, and $R_0 = 5\,\Omega$. These are typical values for R_i, R_o, and A. Using these values, the gain becomes:

$$\frac{v_o}{v_s} = 9.999899906 \tag{8.70}$$

Suppose, however, that we make an approximation that the gain of the op-amp, A, approaches infinity. That is, we evaluate:

$$\frac{v_o}{v_s} \approx \lim_{A\to\infty} \frac{AR_i\left(R_1+R_2\right)+R_2R_0}{AR_2R_i+R_o\left(R_2+R_i\right)+R_1R_2+R_i\left(R_1+R_2\right)} \tag{8.71}$$

$$\frac{v_o}{v_s} \approx \frac{R_1+R_2}{R_2} \tag{8.72}$$

$$\frac{v_o}{v_s} \approx 10 \tag{8.73}$$

The percentage error between our approximation and the exact expression we derived is 0.001%. This is an excellent approximation!

8.6.2 Inverting Amplifier

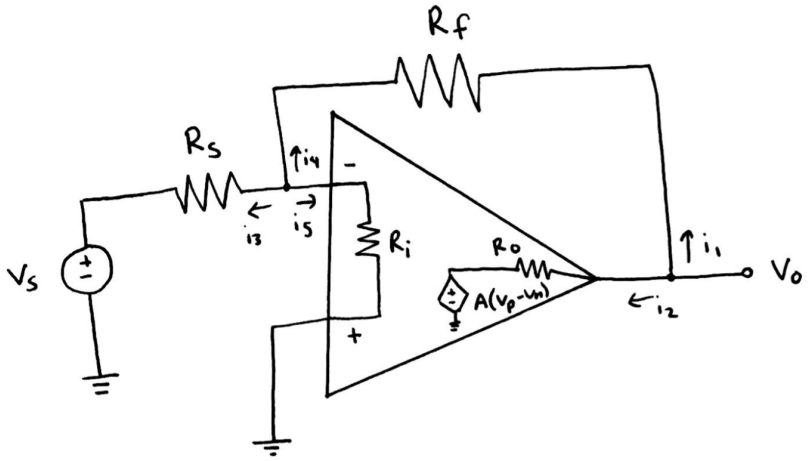


Figure 8.24: An inverting amplifier using the equivalent circuit model of an op-amp

Unlike the noninverting amplifier, the output of the inverting amplifier will have the opposite polarity of the input signal. The expression for the gain of the inverting amplifier can be derived by writing KCL equations at the negative terminal and the output terminal.

KCL @ v_o

$$i_1 + i_2 = 0 \tag{8.74}$$

Next, we substitute the expressions for i_1 and i_2, and we arrive at the following equation:

$$\frac{v_o - v_n}{R_f} + \frac{v_o + A\left(v_p - v_n\right)}{R_o} = 0 \tag{8.75}$$

KCL @ v_n

$$i_3 + i_4 + i_5 = 0 \tag{8.76}$$

Next, we substitute the expressions for i_3, i_4, and i_5, and we arrive at the following equation:

$$\frac{v_n - v_s}{R_s} + \frac{v_n - v_o}{R_f} + \frac{v_n - v_p}{R_i} = 0 \tag{8.77}$$

By using equation (8.75), equation (8.77), the fact $v_p = 0$, and a tremendous amount of algebra, we can derive the expression for the gain of the inverting amplifier as:

$$\frac{v_o}{v_s} = \frac{R_i R_f R_o - A R_i R_f^2}{R_s R_f^2 + R_s R_i R_f + R_i R_f^2 + R_o R_s R_f + R_o R_i R_f + A R_s R_f R_i} \tag{8.78}$$

Suppose that for one particular application, $R_f = 90\,\text{k}\Omega$, $R_s = 10\,\Omega$, $A = 10^6$, $R_i = 10\,\text{M}\Omega$, and $R_0 = 5\,\Omega$. These are typical values for R_i, R_o, and A. Using these values, the gain becomes:

$$\frac{v_o}{v_s} = -8.999990991 \tag{8.79}$$

Let us make the same approximation that the gain of the op-amp, A, approaches infinity. That is, we evaluate:

$$\frac{v_o}{v_s} \approx \lim_{A \to \infty} \frac{R_i R_f R_o - A R_i R_f^2}{R_s R_f^2 + R_s R_i R_f + R_i R_f^2 + R_o R_s R_f + R_o R_i R_f + A R_s R_f R_i} \tag{8.80}$$

$$\frac{v_o}{v_s} \approx -\frac{R_f}{R_s} \tag{8.81}$$

$$\frac{v_o}{v_s} \approx -9 \tag{8.82}$$

The percentage error between our approximation and the exact expression we derived is 0.00001%. This is a fantastic approximation!

8.6.3 Negative Feedback and Linear Dynamic Range

The previous two examples illustrated the concept of *negative feedback*. Feedback refers to the concept of taking part of the output signal and feeding it back or combining it somehow with the input signal. Feedback is called *positive feedback* if the feedback signal increases the magnitude of the input signal, and feedback is called negative feedback if it decreases the magnitude of the input signal. Positive feedback will cause the op-amp to enter into either the positive saturation region or the negative saturation region (this can be useful in some applications, which we will not cover in this chapter). Negative feedback, on the other hand, is essential for controlling the gain and output behavior of the op-amp circuits that we will consider in this chapter.

At this point, negative feedback may seem like a terrible idea. Why would we want to decrease the magnitude of the input signal when we want to amplify it at the output? There are two primary reasons why negative feedback is useful in this context:

1. Although the open loop gain A of the op-amp is very large, on the order of 10^6 or more, the exact value of the open loop gain often varies considerably between different kinds of op-amps and is not a clearly defined quantity. By using negative feedback, we can instead rely on the closed loop gain G which will be a well-defined quantity that that does not depend on A (as long as A is very large).

2. By decreasing the closed loop gain of the op-amp circuit we can improve the *linear dynamic range* of the overall circuit.

Linear dynamic range refers to the range of inputs to a circuit where the outputs will remain in the linear region. As we previously discussed, the op-amp circuit cannot generate an output voltage v_o that is greater than the supply voltage V_{cc}.

$$|v_o| \leq V_{cc} \tag{8.83}$$

In the case of negative feedback, the input and output voltages can be related via the closed loop gain G, thus for the output to remain in the linear region, the input voltages are constrained to:

$$|Gv_s| \leq V_{cc} \tag{8.84}$$

This can be rewritten as:

$$|v_s| \leq \frac{V_{cc}}{G} \tag{8.85}$$

Therefore, we see there exists a trade-off between gain and linear dynamic range. This trade-off can be clearly seen in Figure 8.25.

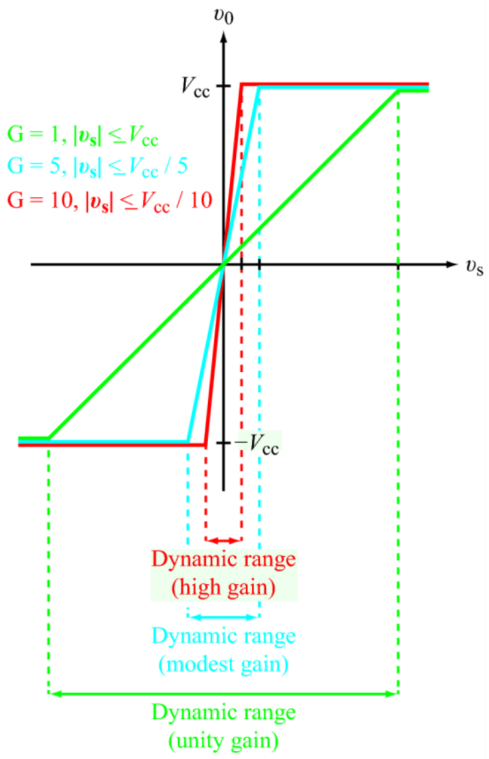


Figure 8.25: Input-ouput transfer characteristics for various gains

Chapter 9

Inductors and Converters

Another energy storage element, and a useful application

9.1 Learning Objectives

- Understand what an inductor is
- Understand the equation: V=L di/dt
- Understand that:
  - Current cannot change instantaneously through an inductor (an inductor tries to keep current constant)
  - An inductor will generates voltage (in either direction) to resist current changes
- Understand that ideal inductors and capacitor are lossless:
  - They store energy and don't dissipate it.
  - Energy that goes into an LC circuit, must come out
  - The energy stored in an inductor is: $\frac{1}{2}L \cdot i_L^2$
- Therefore inductors and capacitors can be used to convert energy from one voltage to another in a theoretically lossless way.
- The size of an inductor or a capacitor is related to the energy they can store.
- Be able to use impedance to:
  - Solve for the output voltage of a buck converter
  - Determine the needed switching freq given L,C (or vice versa)
- Be able to solve for the currents in a Buck or Boost voltage converter

9.2 What are inductors

We briefly introduced inductors informally in the impedance section, but so far haven't used them for anything nor formally described what they are. In this section we will find out what an inductor is, and its important device characteristics.

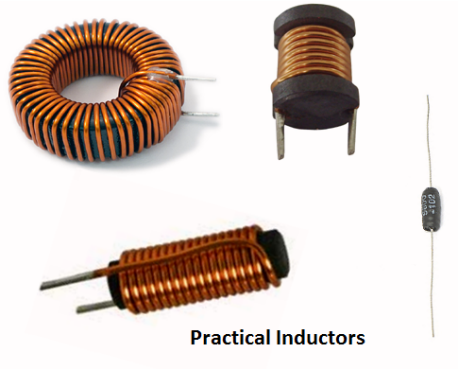


Figure 9.1: Some real inductors

An inductor is another two terminal energy storage device, analogous to the capacitor. The circuit symbol used to represent an inductor is shown here. The value of an inductor is described as its "inductance", usually represented by the symbol "L", and has units of Henries(H). For example, an inductor may have a value of 1H, which is one Henry (though that is a very large inductance, and would be a large physical object. More typical value range from mH to nH).

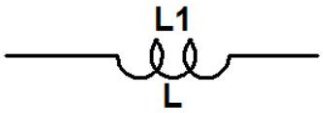


Figure 9.2: Symbol of an inductor

The relationship between voltage and current across an inductor is defined as:

$$V_L = L\frac{di_L}{dt}$$

which can be written as:

$$i_L = \frac{1}{L}\int V_L dt$$

Like a capacitor, the relationship between current and voltage depends on a derivative, but for an inductor it is its voltage that depends on the rate of change in the current, while for a capacitor,

it is the current that depends on the rate of change in the voltage. Also like a capacitor an inductor stores energy, but for an inductor the energy stored depends on the current, and so changing the current requires power to flow either into or out of the device. This means that for an inductor the current can't change rapidly, but the voltage across the device can change quickly, which is exactly opposite the constraints on the current and voltage of a capacitor (voltage can't change rapidly, but current changes can be large. We will discuss where these constraints come from in the next section.

As we can see from the device equations for an inductor, if the current is a sine wave, the output voltage will also be a sinusoidal at the same frequency, but shifted in phase (derivative of a sine wave is a cosine wave) and a magnitude that depends on the value of the inductance, L, and the frequency. The resulting impedance of the inductor is:

$$Z_L = j \cdot 2\pi \cdot f \cdot L$$

Since the Chapter 7 has already work out many example of using the impedance of in building filters, we won't be discussing inductor impedance too much more in this chapter. Instead, we will focus on the time behavior of resistor/inductor circuits, and energy efficient switching power supplies.

9.2.1 Properties of an inductor

As we saw in the previous section, since the voltage across the inductor is:

$$v_L = L \cdot \frac{di_L}{dt}$$

- The current going through an inductor cannot change instantaneously. If it did, then $\frac{di_L}{dt} = \infty$, which gives us $v_L = L \cdot \infty$, which is an impossible phenomenon - we cannot have infinite voltages!
- The voltage across an inductor will change instantly in order to resist any sudden changes in current.

Together these mean that an inductor looks like a current source for short periods of time!

Question 1

An inductor with one terminal connected to ground has no current flowing through it, ie. $i_L = 0A$. At time t=0 switch is turned that connects the other end of the inductor to a 50V power supply. What is the current flowing through the inductor immediately after the switch is turned on?[1]

Question 2:

A different inductor with one side connect to ground is connected in parallel with a 100 ohm resistor. Initially the inductor has a constant 500mA flowing through it, supplied by an external current source. At time t=0, the source providing the current is removed. What is the current through the inductor immediately after time t=0? What is the voltage across the inductor a this point?[2]

[1] Q1 - The current flowing through the inductor immediately after the switch closes is still nothing (0A), since the current can't change instantly. The voltage across the inductor becomes 50V, and the current starts to ramp up.

[2] Q2 - The current flowing through the inductor immediately after time t-0 is still 500mA, since the current can't change instantaneously. To satisfy KCL, this current must come from the resistor, so the voltage across the inductor jumps to -50V, and the current through the inductor begins to ramp down. At this voltage the current through the resistor matches the current flowing into the inductor.

These question show that an inductor initially looks like a current source, and then start changing their current. Since the rate of change in current depends on V/L, if the inductor is large enough the inductor can be current source like for a significant period of time. During this period the inductor uses its ability to store energy to change its voltage, supplying or absorbing energy to keep the current constant. This is similar to a capacitor's ability to change its current to keep the voltage constant.

If you have taken electricity and magnetism in physics, it is possible to get a better understanding of the physical origins of inductance. An inductor is generally just a wire, often wrapped around some other material. This is clearly shown in Figure 9.1. Current flowing in a wire creates a magnetic field. When you wrap wire into a coil the field from all the wires add together to create a larger field. Thus the magnetic field you create is proportional to N the number of loops of wire in the inductor. The material that the wire is wrapped around generally has a high magnetic permeability μ, which further increases the magnetic field. Since creating a magnetic field requires energy, this is how an inductor stores energy. The voltage generated by an inductor comes from Faraday's Law, which states that the voltage generated around any wire loop is proportional to the change in magnetic field times the area (the magnetic flux) that the loop sees. Since all the wire loops in an inductor are connected in series, the voltage from all these loops add up. The resulting voltage produced is proportional to:

$$V_L = N \cdot A \cdot \frac{dB}{dt} = N \cdot A \cdot N \cdot k\mu \frac{di_L}{dt} = k\mu \cdot A \cdot N^2 \frac{di_L}{dt}$$

where N is the number of turns, μ is the magnetic permeability of the core, and A is the area of the wire loop. This is where our equation describing the voltage and current relationship in an inductor comes from, and shows why high value inductance generally has many turns of wire in it. Faraday's law says that as current increases through a coil of wire, a magnetic field is generated, creating a voltage opposing that change in current.

9.2.2 Energy stored in an inductor

Remember that energy is just the integral of power over time, and that the power flow through a device is always $P = I \cdot V$. This gives:

$$v_L = L \cdot \frac{di_L}{dt}$$

$$P = i_L \cdot v_L = i_L \cdot L \cdot \frac{di_L}{dt}$$

Integrating this equation for the power flowing into an inductor yields the energy stored into the inductor.

$$E_L = \int_0^t P dt = \int_0^t i_L \cdot L \cdot \frac{di_L}{dt} dt$$

Notice that the right hand side of the equation no longer depends on time. Lets assume that the current in the inductor started at 0, and at the end, time = t, the inductor current is i_{final}. This gives the energy stored into the inductor as:

$$\int_0^{i_{final}} i_L \cdot L \cdot di_L = \frac{1}{2} \cdot L i_{final}^2$$

9.2.3 Transformers

Transformers are basically two inductors "coupled" together which allows us to convert voltages. As shown in Figure 9.3, the two inductors are coupled through an iron core, which provides a very low loss path for magnetic field (or in the case of an ideal transformer, a no loss path) such that all of the field that travels through one inductor also goes through the second.

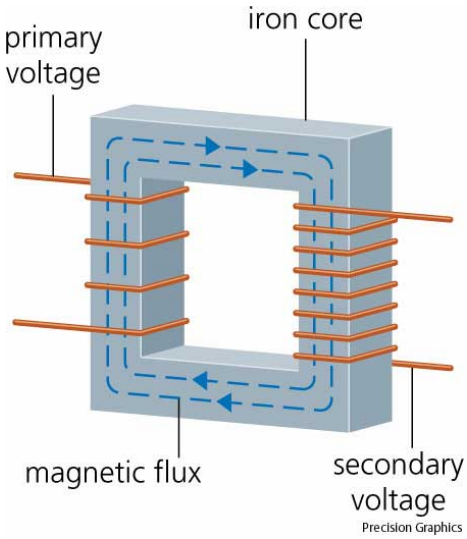


Figure 9.3: How a transformer works

Transformers are used for two functions. The first is isolation. The first coil converts the energy that was driven into its terminals into energy in the magnetic field that is flowing through the transformer. The second coil that then extract energy from this magnetic field, without having any direct electrical connection to the circuit driving the first coil. In addition to isolation, transformers can also change the amplitude of input signal if the number of turns in the two coils is not the same. Remember that the transformer is sharing the magnetic field, and the conversion from changing magnetic field to voltage depends on the number of wire turns. Thus if you want to convert a high voltage, say 120V AC into a smaller voltage, like 6V, and you want the output voltage to be isolated from the 120V supply, a transformer is the perfect device to use. By having the input coil have 20 turns (120/6) for each turn in the output coil, a transformer can both reduce the voltage and provide isolation between the two circuits. These devices are used in almost all the power converters that you use today.

If a transformer has a different turns ratio, it transforms the current as well as the voltage, but the current is transformed in the opposite direction. Assume we are still working with our 120V to 6V transformer. We can now ask how much current must we put into the 6V end to cancel the magnetic field generated by the 120V end. Since the size of field depends on the current and then number of turns, if I put in 1mA of current in on the 120V coil, I would need 20x more current, or 20mA on the 6V end to cancel the field. This results makes sense, since with this current and voltage ratio energy is still conserved. That is the power flowing into the 120V end is equal to the power flowing out of the 6V end.

9.2.4 Ideal inductors vs. real inductors

So far we have been discussing inductors as ideal components which have no loss. In reality, of course, nothing is ideal, and almost all inductors have loss.[3] As we have seen inductors are made out of wire, and although we typically assume wires are loss less and have no resistance, we know the truth is that they do possess **some** resistance. The resistance of a wire is proportional to its length, and inversely proportional to its area. Hence short fat wires have very low resistance, while long skinny wires have larger resistance. This often poses a problem for creating higher value, low loss inductors, since for these inductors we need to generate a coil with many turns. But this is only possible if the wire is not too large, since each time the wire is wrapped around a core it occupies at least its wire diameter. Getting higher inductance at higher current generally requires larger, more expensive inductors.

We model a real inductor as the series combination of two idea elements, an ideal inductor with no loss, and a resistor. The resistor models the loss from current flowing through the wire, and the inductance models how changing the magnetic field changes the voltages in the circuit.

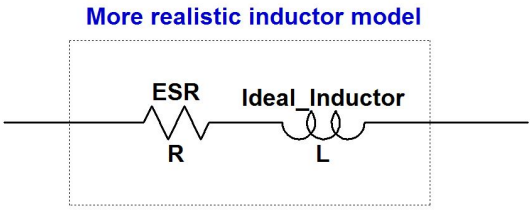


Figure 9.4: Modeling a real inductor

This means that when current flows through the inductor, some of the voltage drop falls across that resistance. Hence, the effective voltage across the "ideal inductor" in this model is less than what we see across the inductor. Therefore, the same voltage across the inductor will correspond to less of a change in current than it would for an ideal inductor. In other words, the model tells us that the real inductor no longer stores as much energy as it used to, since it now has some loss.

The model also allows us to calculate the amount of real power lost in a real inductor by calculating the loss of the ESR (equivalent series resistance).

[3]To have no loss an inductor would need to use wire that has no resistance. While this might seem impossible, there are materials called superconductors, where their resistance drops to 0 below a certain temperature. Since this usually only happens at very cold temperatures we will not talk about them further in this book.

9.3 LR Circuits

Using these concepts about the properties of inductors, we can solve for the voltages and currents in some simple circuits with inductors and resistors.

Example 1: L-R circuit driven by a step input

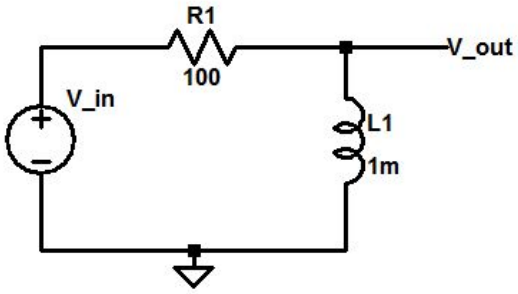


Figure 9.5: L-R circuits example 1 circuit

Let's examine what happens if the voltage source on the left changes its voltage from 0V to 1V. Thinking qualitatively first will help understand how the circuit works. Before the voltage rises, V_{in} is ground, and no current can flow through any device. Thus the current through the resistor and inductor are both zero. Hence, immediately after the step occurs, $i_L = 0$ still, because current through an inductor cannot change instantaneously. Since the inductor is in series with the resistor, $i_R = 0$ at this time also. There is therefore initially no voltage drop across the resistor, and all of the voltage from the input source appears at the output and $V_{out} = V_{in}$. This step in the output voltage causes a voltage across the inductor, which causes the current to ramp up, which causes the current through the resistor to ramp up as well. The voltage drop across the resistor, caused by the growing resistor current lowers the output voltage, which slows the change in the output current. This is the same situation we saw with a capacitor where the rate of change in the output was proportional to the output which led to an exponential waveform. Eventually we reach a point where the voltage across the resistor is 1V, and the output voltage (and the inductor voltage) is 0V. At this point the current stops changing, since the inductor voltage is zero, and we say that the circuit has reached steady state. This means that the inductor basically becomes a piece of wire connecting the resistor to GND, or 0V, and the current through the resistor is simply $i_R = \frac{V_{in}}{R} = \frac{1}{100} = 10mA$.

These are the plots of the voltage and current waveforms in response to the step input. Does the output voltage correlate with our qualitative assessment? Does the current?

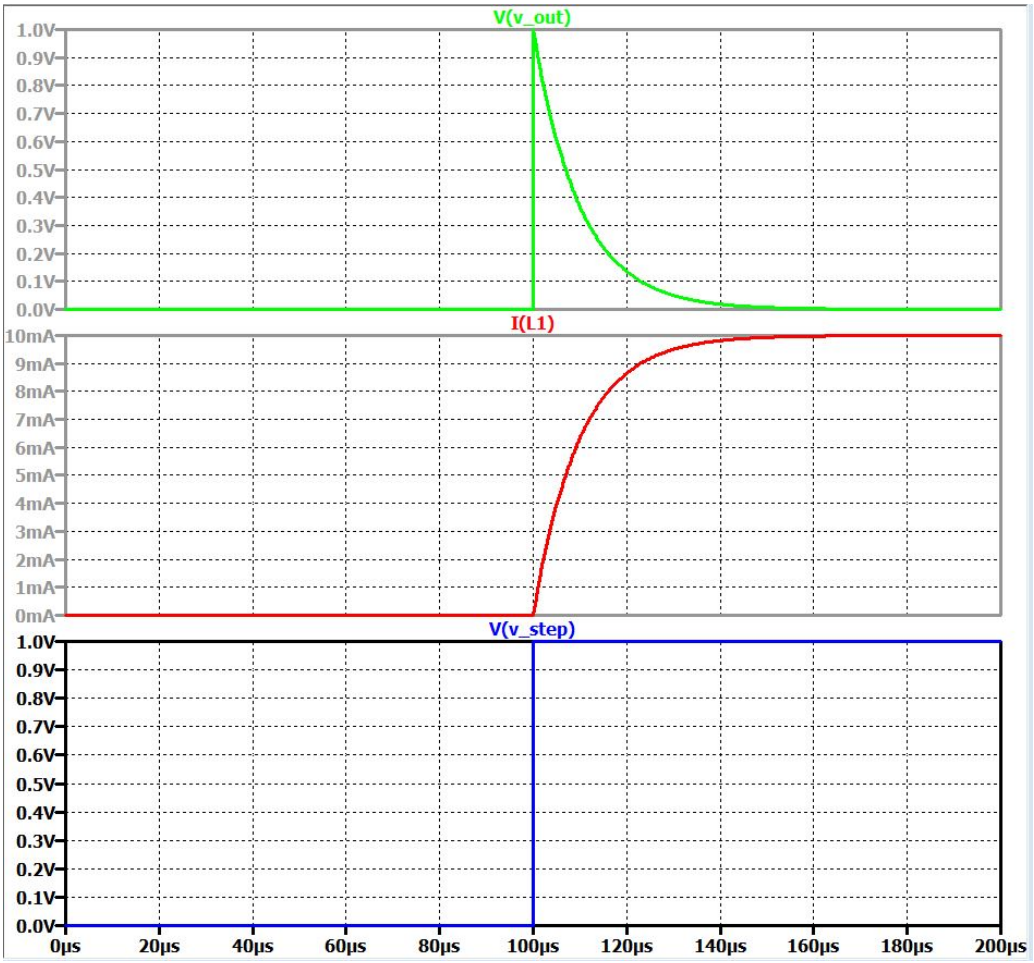


Figure 9.6: L-R circuits example 1 waveforms in time domain

Example2: L-R circuits driven by a step input

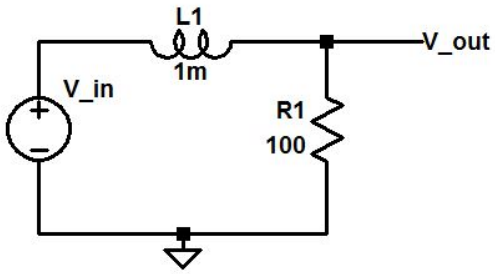


Figure 9.7: L-R circuits example 2 circuit

As an exercise, try to qualitatively determine the current and voltage through and across the inductor when the step happens, and a long time after the step happens (at steady state).

So again, $i_L = 0$ and $i_R = 0$ to start with. However, this time the resistor is connected between V_{out} and GND, and since at this point in time the voltage across the resistor is 0, the output voltage starts at 0V instead. Once the circuit reaches steady state, the voltage across the inductor becomes 0V, and the output voltage rises to V_{in}. The correct current and voltage waveforms are shown in Figure 9.8.

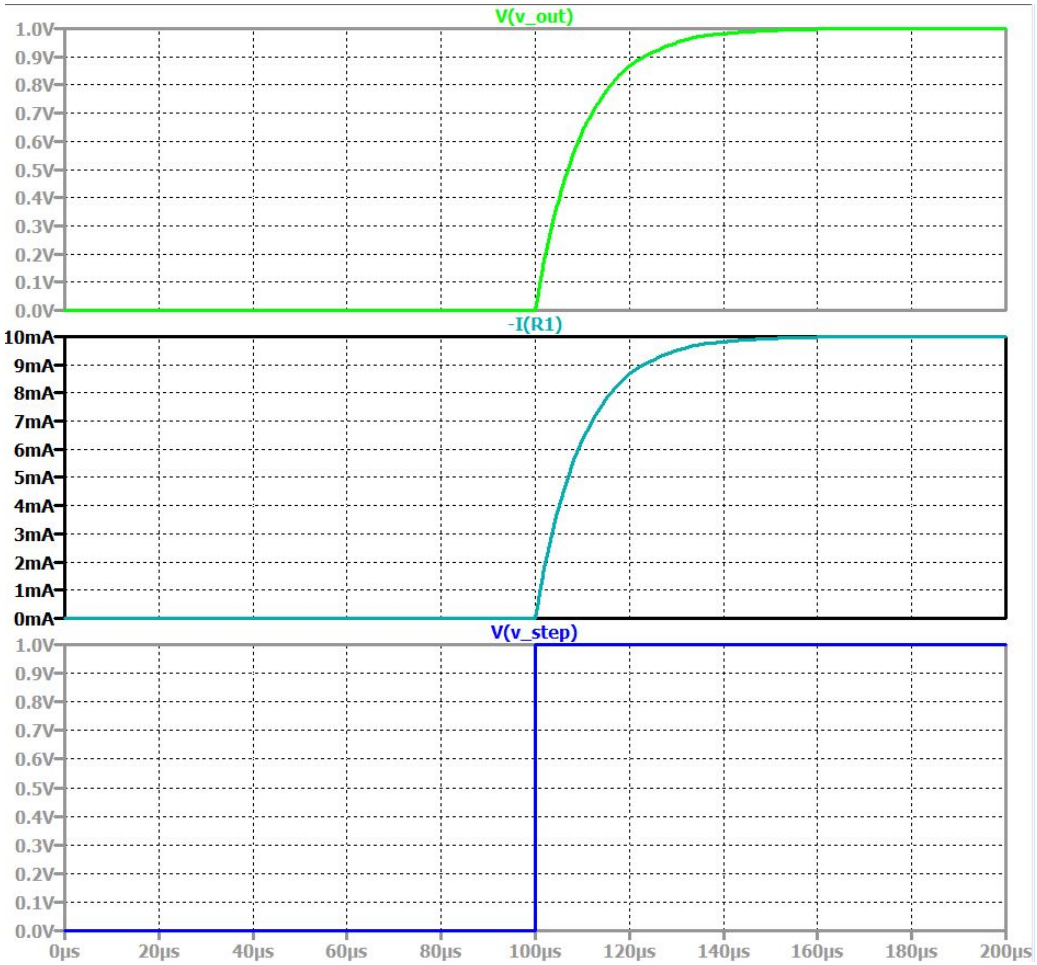


Figure 9.8: L-R circuits example 2 waveforms in time domain

Example 3: An L-R circuit with a switch Let us consider a slightly more complex circuit, as shown in Figure 9.9. When the switch has been on for a long time, the inductor looks like a short circuit, and all of the current flows through the inductor and through R2. This current will be $\frac{1V}{10\,\Omega} = 0.1A$. If the inductor is ideal, no current will flow through R1.

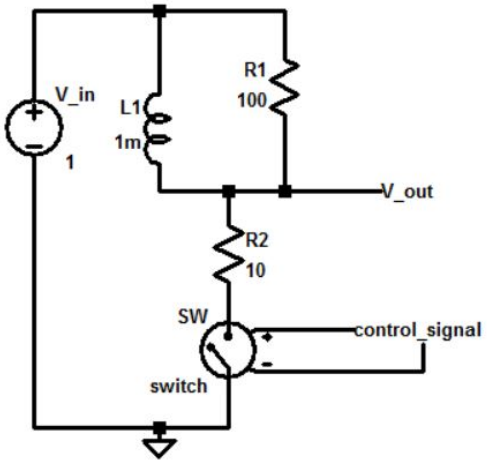


Figure 9.9: L-R circuit with a switch

What happens when we disconnect the resistor from ground by turning off the switch? Again we need to start with the basic rules for an inductor: its current can't change instantly, so when the switch is off, it will have the same current it had the moment before, 0.1A, and this current will be flowing from the 1V supply and out the end of the inductor that is no longer connected to GND. Since this current can't flow through R2 (it would violate KCL at the switch node) it must take another path. The only path possible is to flow through R1, and put the current back into the 1V power supply.

But the voltage across that resistor was initially 0V, and no current was flowing through the resistor. To get 0.1A to flow, the resistor needs $V = iR = 10V$ across it, and so that is what the inductor does, it drives V_{out} to 11V, 10V above the supply to support the inductor current.

This voltage then begins to decrease the current in the inductor which decreases the output voltage, until the output settles down ack at 1V.

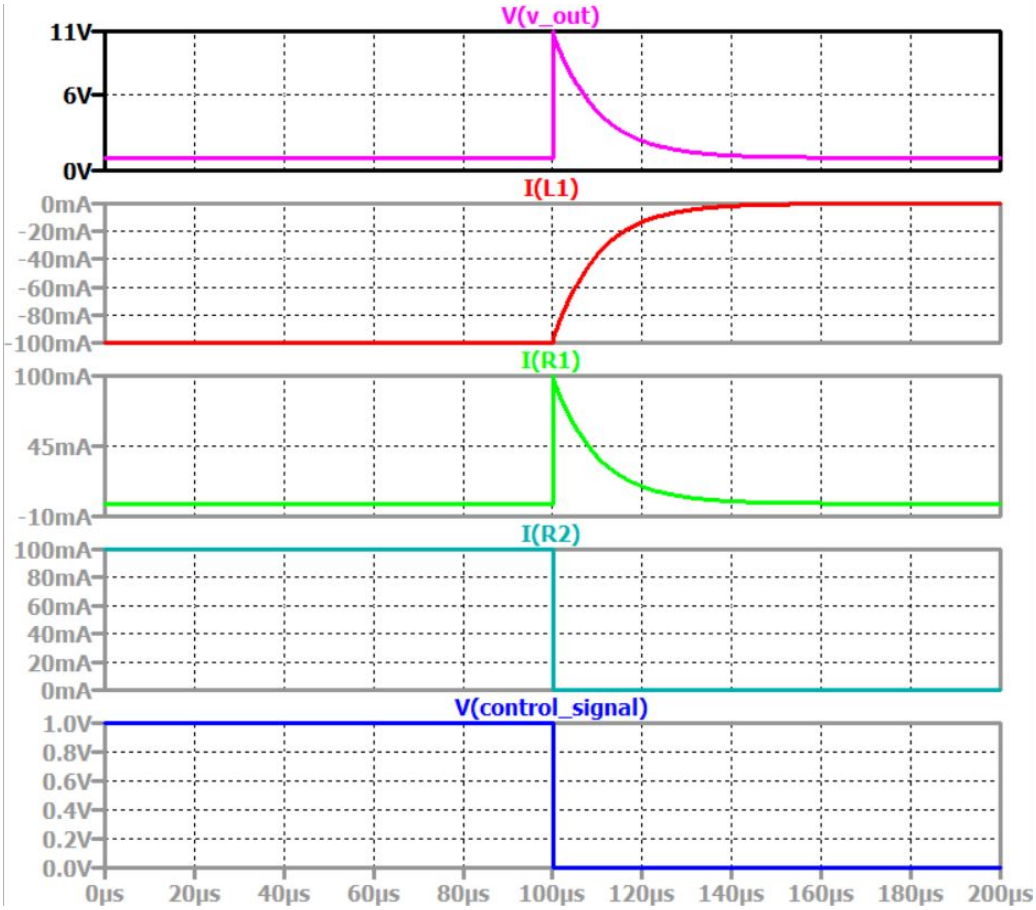


Figure 9.10: L-R circuit with a switch - waveforms

9.4 Switching power supplies

We have seen a couple of times in the class where we control the power we provide to an object by switching between turning it on at maximum power, and not providing power at all. This type of control is used in your soldering irons to control the tip temperature, your air-conditioner, refrigerator, oven, and electric stove. You might have used this method to control your motor speed, or your LED brightness. The reason this method of control is so popular is it's very energy efficient. The circuit can control kiloWatts of power, and only require 10s of Watts to operate. The key to this efficiency is that either the controller provide no energy (which doesn't have any loss), or just needs to connect the device to the power supply. If the resistance of the switch making this connection is small, this too should be low loss.

This kind of on-off switching is a good control strategy when the system you are building has something that can filter this energy to create the desired average. This happens automatically in systems that deal with temperature, since the temperature is essentially the integral of the energy

that you put in or take out. That is why most heating/cooling systems us on off control. We want to take this same idea, and apply it to change the voltage of an energy source. Since in this case we are putting in voltage and want to get out voltage, there is no intrinsic filtering in this system. But we can use inductors and capacitors to create an electronic filter, and this will allow us to efficiently create any voltage we need with very high efficiency. This section will explain how two different converters work, a Buck converter which generates an output voltage lower than the input, and a Boost converter which generates an output voltage that is larger than the input voltage. Before going through the detailed operation of a Buck converter, we will first explain how it works using a simple filter model.

9.5 Buck Converter

A buck converter comprises essentially an inverter driving a LC filter. The circuit is shown in Figure 9.11.

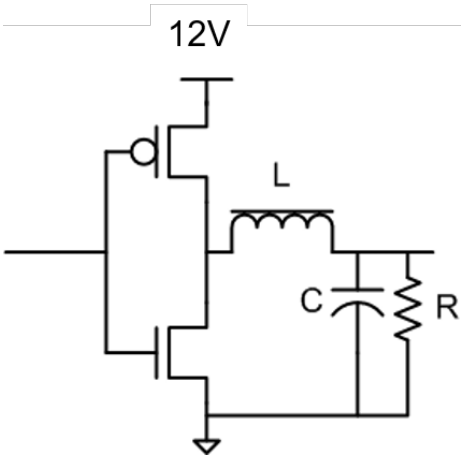


Figure 9.11: Buck converter circuit

The inverter is the on-off control in this circuit. When the inverter input is low, the pMOS transistor connects the output to the power supply, which supplies voltage and energy to the LC filter and the output. When the input is high, the nMOS connects the output to GND, driving the output to GND and no energy is provided to the filter or output. The point of the pMOS and nMOS transistor is to provide very low resistance paths for the current going to the load, so these devices are often power transistors, with very low on-resistance. If the resistance of both of these transistors are small, then there will be little power dissipated in the converter. While there might be large current flowing through the inductor, the voltage drop across the transistors will be small, and thus the power dissipated by the transistors, iV, will also be small.

One possible inverter output waveform is shown in Figure 9.12. In this example, the duty cycle is 25%. We can work out the relationship between the input duty cycle (25% in this case) and the voltage at the output of the converter using simple frequency domain analysis, and breaking down the inverter output waveform into its frequency domain components (the sine waves you need to

add together to generate the waveform). For any repetitive signal like this one, to break the signal into its frequency components you have to find the principle frequency of the signal, which is the frequency that the signal repeats. This is easily calculated as one over the time it takes the signal to get back to the same point in its pattern. The time for the signal to repeat is called the signal's cycle time. Once we have found the cycle time and thus the primary frequency, f_s, the only possible sinusoid's that can be present in that signal are at $n \cdot f_s$, where $0 \leq N < N_{max}$.

While the frequency components for $N > 0$ are clear, the $N = 0$ component represents the constant value you need to add to the sinusoid to match the waveform. This value is easy to calculate, since the average value of any sinusoid is zero: adding sinusoids doesn't change this average value. If the average value of your waveform is not zero, we need to explicitly add this average value in to our sinusoids to make the waveform match. Since this value doesn't change with time, we call this component the DC (direct current) or zero frequency component. For the waveform shown in this figure the average value will simply be $25\% \cdot VDD$.

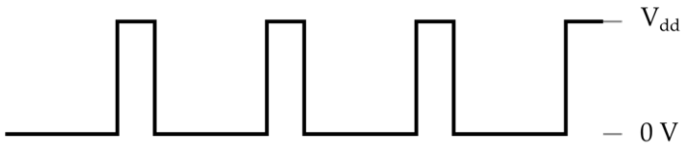


Figure 9.12: Square wave input to filter created by inverter

Given the understanding of how to convert the output of the inverter into frequency components, we can easily estimate what the output of the converter should look like using the impedance of resistors and capacitors. The filter network of a buck converter is shown in Figure 9.13.

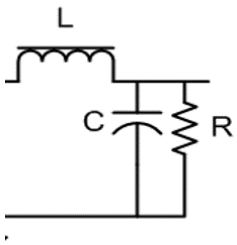


Figure 9.13: The filter network of the buck converter

The transfer function of this filter can be found by first identifying that R and C are in parallel. We can call their combined impedance Z2. Then the filter is just an impedance divider, and the transfer function can be found as follows:

$$TF = \frac{Z_2}{Z_1 + Z_2}$$

$$Z1 = Z_L = 2\pi \cdot f \cdot L$$

$$Z2 = Z_R \parallel Z_C = R \parallel \frac{1}{2\pi \cdot f \cdot C} = \frac{R}{1 + 2\pi \cdot f \cdot R \cdot C}$$

Therefore

$$TF = \frac{\frac{R}{1+2\pi \cdot f \cdot R \cdot C}}{2\pi \cdot f \cdot L + \frac{R}{1+2\pi \cdot f \cdot R \cdot C}}$$

$$TF = \frac{1}{1 + 2\pi \cdot f \cdot \frac{L}{R} + L \cdot C \cdot (2\pi \cdot f)^2}$$

The Bode plot of this filter will look something like Figure 9.14 - it actually depends on the values of L, R, and C, but in understanding the operation of the buck converter we will consider this specific example.

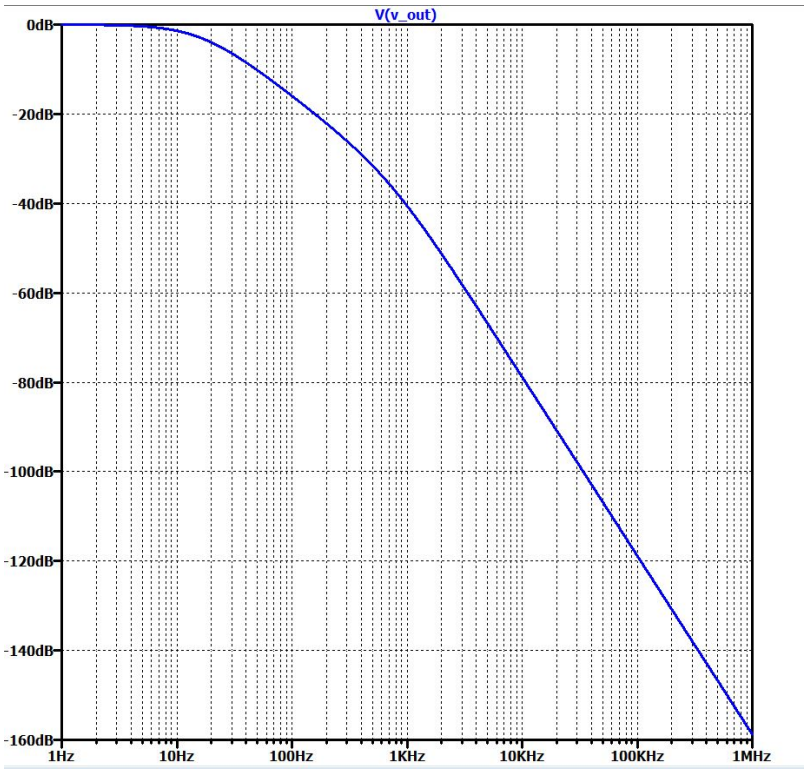


Figure 9.14: The filter network of the buck converter

Given an ability to calculate how the filter attenuates different frequency components, and some understanding of how to generate those frequency components from the input waveform, we can estimate the output waveform of a real converter.

Question: If the inverter input repeats its pattern every 10 µs what is the lowest frequency sinusoidal component in the output?[4]

If the output filter greatly attenuates this frequency and the any higher frequency component, then the only signal that will remain is the zero frequency component of the input, or the average value of the input waveform. This is a great result if it is true, since we can then change the output voltage by just changing the duty cycle of the input waveform (which changes the average value of the output waveform).

Example:
The input voltage (V_{in}) to a buck converter is 12V and we want a 6V output voltage, (V_{out}). What duty cycle, D, will yield the correct output voltage?

$$V_{out} = D \cdot V_{in}$$

$$D = \frac{V_{out}}{V_{in}} = \frac{6}{12} = 0.5 = 50\%$$

Therefore the inverter should be operated at 50% duty cycle.

Let's assume we are using a 40 µH inductor and a 600 µF capacitor, and the load we are driving is effectively a 6 Ωresistor. In this case we can't solve for the residual signal on the output, since we don't know the amplitude of all the different frequency components, but we can estimate an upper bound on the signal. We start by finding the gain of the filter at the fundamental frequency of 100 kHz. At this frequency the gain will be:

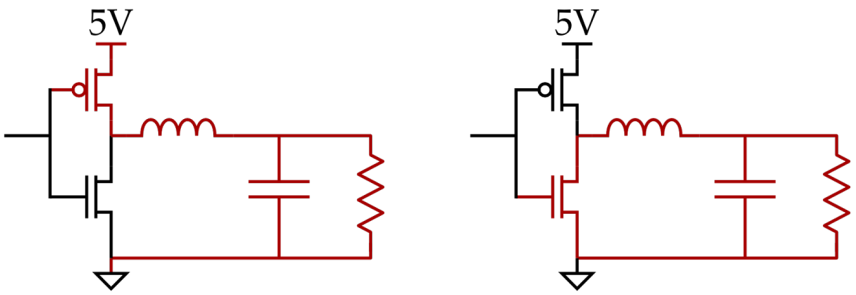
$$Gain = \frac{1}{1 + 2\pi \cdot 10^5 \text{Hz} \cdot \frac{40\,\mu\text{H}}{6\,\Omega} + 40\,\mu\text{H} \cdot 600\,\mu\text{F} \cdot (2\pi \cdot 10^5 \text{Hz})^2} = 7 \cdot 10^{-4}$$

This the gain at the fundamental, f_s, and the gain at $2f_s$ will be 4 times smaller, since the gain is falling off as f^{-2} at this frequency. Since the attenuation at the higher frequencies will be larger, if we attenuate all the sinusoids by the attenuation of the fundamental tone, we will over estimate the size of the output signal. But this is ok, since it will create a max on the size of the ripple, and is easy to calculate. For our 50% duty cycle signal the average value is 6V, and the sum of all the sinusoids is a waveform that goes up and down 6V. If this signal is multiplied by a gain of $7 \cdot 10^{-4}$, the resulting signal will be a signal that is only $6V \cdot 7 \cdot 10^{-4} = 4\,\text{mV}$ which is very small compared to the 6V output voltage.

9.5.1 Detailed Analysis of a Buck Converter

The buck converter essentially has two states - either the top switch is on, or the bottom switch is on. Figures 9.15a and 9.15b show how the current flows through the buck converter in the two different states of the buck converter.

[4] The lowest frequency sinusoidal component of the output will be equal to the switching frequency, which is $1/10\,\mu\text{s} = 100\,\text{kHz}$



(a) Current flow through buck converter (1)

(b) Current flow through buck converter (2)

Figure 9.15: Operation of a buck converter

We could attempt to solve each state of the circuit using KCL and KVL, but since there is both an inductor and a capacitor, we will end up with some second order differential equations. A simpler method to solve them would be preferred, and this can be achieved by making two approximations:

- The output voltage is approximately constant.
- The circuit will settle into some repeating cycle - ie. The voltage across the capacitor, v_c, and the current through the inductor, i_L return at the end of every cycle to the same value they had at the beginning of that cycle.

The first approximation can be made because we are designing the circuit to reach that particular objective - we want a stable output voltage. The second approximation is made because we believe that the output will eventually settle down and become periodic, like the input waveform that is driving it. If the output voltage is periodic, it must return to the same voltage on each cycle. Using only the first assumption, that the output is constant, it follows that the inductor only ever has two different voltages across it:

$$5\text{V - } V_{out} \text{ in state (1), and } -V_{out} \text{ in state (2).}$$

Since $v_L = L \cdot \frac{di_L}{dt}$, $\frac{di_L}{dt} = \frac{v_L}{L}$ the inductor current will ramp up at a rate of:

$$\frac{5V - V_{out}}{L} \text{ in state (1), and } \frac{-V_{out}}{L} \text{ in state (2).}$$

If the output waveform is going to be periodic, the inductor current must return to the same value in each cycle (this is our second assumption). These assumptions are visualized in Figure 9.16 which shows the inductor voltage and current in the converter. This figure also shows the input waveform to the inverter which drives the inductor, so when the input waveform is low, the pMOS transistor is connecting the inductor to Vdd, which increases its current.

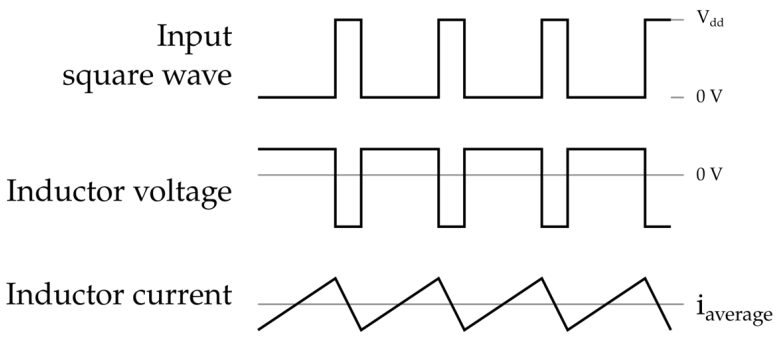


Figure 9.16: Buck converter - circuit waveforms

We will define the time during which the input square wave is low, so the inductor is connected to VDD to be t1, and the time which the input square wave is high so the inductor is driven to GND to be t2, as shown in Figure 9.17,

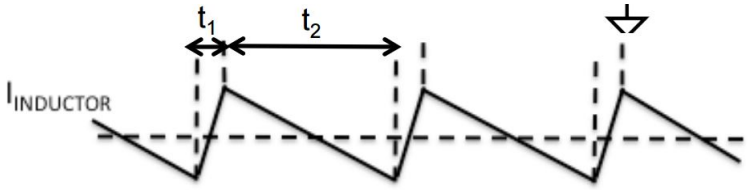


Figure 9.17: Buck converter inductor current waveforms

The following equations can then be written to describe how much the current charges up during t1, and how much it discharges during t2:

$$\Delta i_{charge} = t1 \cdot \frac{5V - V_{out}}{L}$$
$$\Delta i_{discharge} = t2 \cdot \frac{-V_{out}}{L}$$

Since the current at the start and end of the cycle are the same the net change in current in the inductor should be zero:

$$\Delta i_{charge} + \Delta i_{discharge} = 0$$
$$t1 \cdot \frac{5V - V_{out}}{L} + t2 \cdot \frac{-V_{out}}{L} = 0$$

Rearranging,

$$V_{out} = 5\,\mathrm{V} \cdot \frac{t1}{t1 + t2}$$

And since duty cycle can be described as $D = \frac{t1}{t1+t2}$, we arrive at the same equation we found earlier:

$$V_{out} = 5\,\mathrm{V} \cdot D$$

Again, it is found that the output voltage of the buck converter doesn't depend on load current (i_R) at all! In fact as we will see in the next section, it doesn't even depend on the direction of the current flow through the inductor.

9.6 Boost Converter

Figure 9.18: The buck converter PCB board used in the solar charger lab

Figure 9.18 is a photograph of the boost converter you used in your first lab, building the solar charger. This converter needed to create an output voltage that was higher than its input voltage, producing a 5V output from a 3-4V input voltage. At first this seems hard. How can one create a voltage larger than the voltage you start with. But inductors can do that easily. In fact, Figure 9.9 does exactly that. It creates an 11V output spike from a 1V power supply. We use the same basic technique, charge up an inductor and then use the inductor to drive the higher voltage output.

What is most amazing about Boost converters, is that we have essentially already studied them: they are normal buck converters that we run the energy flow backward. That is we connect our lower voltage energy source to the output node the the buck converter, and the higher voltage output is what we used to call the input port. This is shown in Figure 9.19

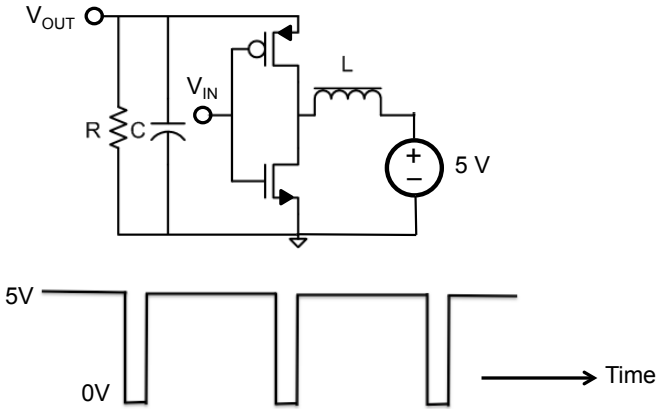


Figure 9.19: Schematic of a Boost converter, with its input waveform

The current balance equations are the same form that there were in the Buck converter, but the value of the sources have changed. The 5V source in the buck converter is now Vout, and what was Vout in the Buck converter is now the input voltage, which we will call Vsupply (5V in the figure). Since current nominally flows from the source supply to Vout, we will change the reference direction for the inductor current. Positive current through the inductor flows to the left. Again we assume that the inductor is connected to the pMOS device during T1, and the nMOS device during T2. This give the following equations for the current change in the inductor:

$$\Delta i_{discharge} = t1 \cdot \frac{V_{supply} - V_{out}}{L}$$

$$\Delta i_{charge} = t2 \cdot \frac{V_{supply}}{L}$$

Since the current at the start and end of the cycle are the same the net change in current in the inductor should be zero:

$$\Delta i_{charge} + \Delta i_{discharge} = 0$$

$$t1 \cdot \frac{V_{supply} - V_{out}}{L} + t2 \cdot \frac{V_{supply}}{L} = 0$$

Rearranging,

$$V_{supply} = V_{out} \cdot \frac{t1}{t1 + t2}$$

Or,

$$V_{out} = V_{supply} \cdot \frac{t1 + t2}{t1}$$

And since duty cycle can be described as $D = \frac{t1}{t1+t2}$, we arrive at

$$V_{out} = \frac{V_{supply}}{D}$$