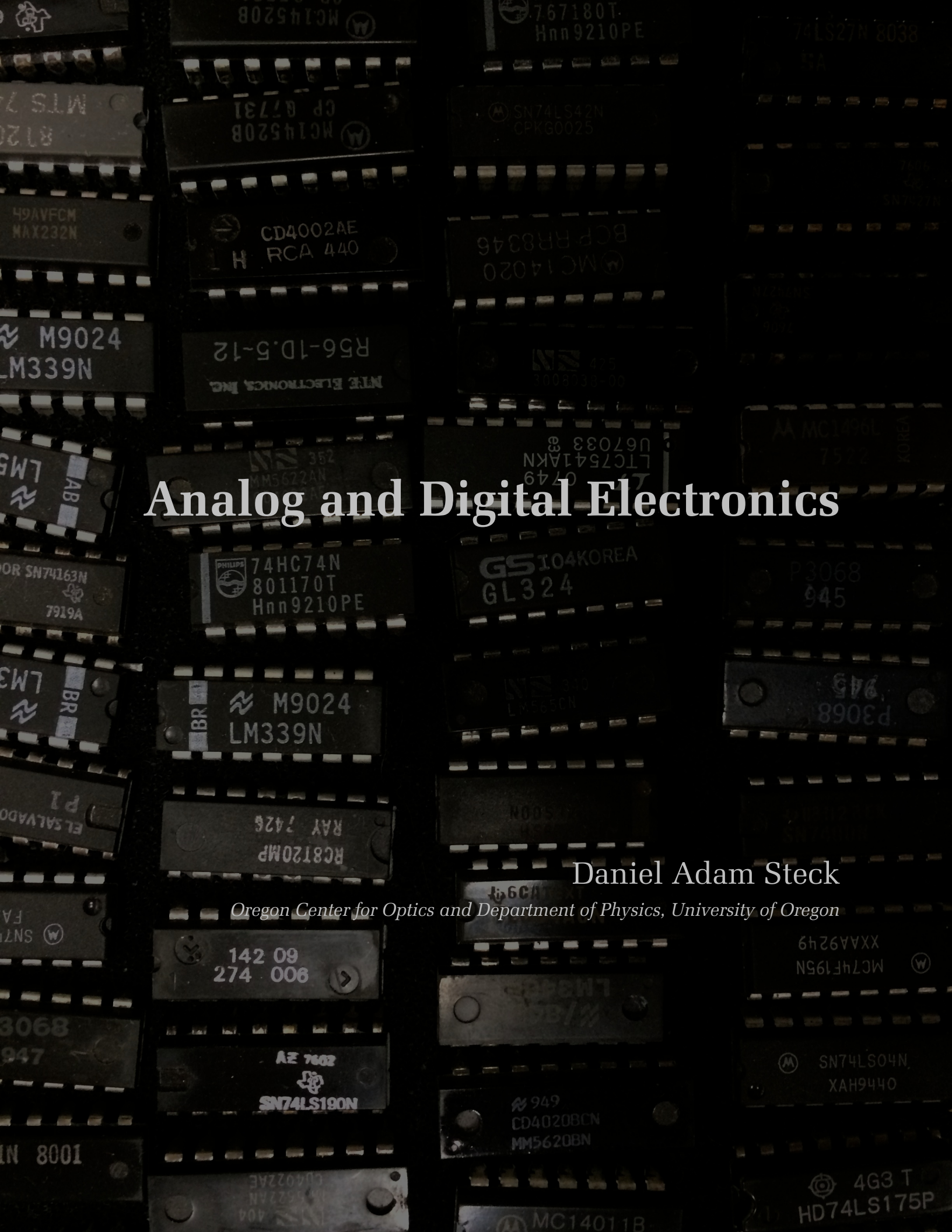


Analog and Digital Electronics

Daniel Adam Steck

Oregon Center for Optics and Department of Physics, University of Oregon

Analog and Digital Electronics

Daniel Adam Steck

Oregon Center for Optics and Department of Physics, University of Oregon

Original revision posted 28 January 2015.

This is revision 0.2.0, 28 March 2017.

Cite this document as:
Daniel A. Steck, *Analog and Digital Electronics*, available online at `http://steck.us/teaching` (revision 0.2.0, 28 March 2017).

Author contact information:
Daniel Steck
Department of Physics
1274 University of Oregon
Eugene, Oregon 97403-1274
dsteck@uoregon.edu

Contents

Part I

Analog Electronics

Chapter 1

Resistors

1.1 Basic Definitions

Here, we're going to breeze through a few fundamental notions in electromagnetism. At the most basic level, electronics studies the flow of "stuff," or more specifically, **charge** (measured in **Coulombs**). The *flow* of charge is **current** (*not* "amperage"), defined by

$$I = \frac{dQ}{dt}. \tag{1.1}$$
(current)

What causes charge to move around and form currents? It's the **potential** associated with an **electric field**. To move a charge between two points, say A and B, this requires some work (energy) W done against the force due to the field. Then the **potential difference** or **voltage difference** is

$$V_{AB} = \frac{W}{Q}. \tag{1.2}$$

That is, the work is proportional to the charge and to the difference in potential between the two points:

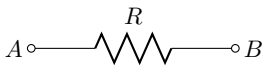
$$V_{AB} := V_A - V_B. \tag{1.3}$$

For a *static* electric field, it turns out that V_{AB} is independent of the path that the charge takes between the points, so we can represent this as a simple difference between the endpoint potentials. It's important to note that only *differences* in potential matter: if we raise both V_A and V_B by the same amount, the work to transport the charge isn't affected. Finally, note that voltage/potential is measured in **volts (V)**, which is the same as joules per coulomb (J/C), as we can see from the work relation (1.2).

An **electromotive force (EMF)** is a special name for a voltage difference due to an energy source (say, a battery, or a power supply).

1.2 Ohm's Law

Since electric fields exert forces on charges, you might think that a constant electric field makes a charge move *ballistically*, or like a mass moves under constant gravity. But charges (electrons) moving through a material (metal, semiconductor), due to interactions with the material, quickly settle into a **terminal velocity**, like a particle falling through air under gravity. Under these conditions, and for small voltages, the velocity of the charges is proportional to the applied voltage, so the *current* is proportional to the voltage. This is the content of **Ohm's law**. Consider a *resistor* (essentially any conducting material, say a wire, where the material "resists" the flow of charge), which we represent by the following schematic symbol:



Here the **resistance** (measured in **ohms** or Ω) is R, and the resistor connects points A and B. Then Ohm's law states

$$V_{AB} = IR. \tag{1.4}$$
(Ohm's law)

That is, for a fixed voltage, the current is inversely proportional to the resistance, which is sensible.

The voltage (and hence, electric field) does work on the charges. The **power** is the rate of work, or

$$P = \frac{dW}{dt}. \tag{1.5}$$

From Eq. (1.2), $W = VQ$ (dropping subscripts on V), and at constant voltage,

$$P = V\frac{dQ}{dt} = VI. \tag{1.6}$$
(electrical power)

Of course, with Ohm's law, we can also write

$$P = VI = I^2R = \frac{V^2}{R} \tag{1.7}$$
(electrical power)

for a few useful alternate forms of the electrical power

1.2.1 Resistors

Essentially anything short of a superconductor has resistance. Wires that carry current have resistance, but usually it's desirable to keep their resistance small. But in virtually all electronic circuits, it's useful to introduce controlled quantities of resistance, and these are the electrical components we call resistors. A few basic types are:

1. **wirewound resistors**: are just wires wrapped around a form. These are usually expensive, but can be precise (using thin wire to make a large resistor) or able to handle high power (using thick wire embedded in ceramic).

2. **carbon film**: are a thin layer of carbon deposited on some insulating form (usually a small cylinder). They're cheap, but not particularly accurate in value. In the type with axial leads, these are usually recognizable by their tan color.

3. **metal film**: are a thin layer of metal deposited on some insulating form (again, usually a small cylinder). They're more expensive than carbon, but more accurate. In the type with axial leads, these are usually recognizable by a blue or blue/green color.

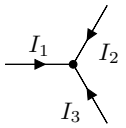
1.3 Networks and Kirchoff's Laws

A simple **circuit** is any network of resistors, batteries, and wires (later to include more stuff!). There are two basic laws that govern the circuit if all we have is batteries and resistors, and these are called **Kirchoff's laws**, on for current, and one for voltage.

1. The **current law** states that at any junction, the current going in to the junction must exactly balance the current going out, for charge not to accumulate there:

$$\sum_j I_{\text{in},j} = \sum_j I_{\text{out},j}. \tag{1.8}$$

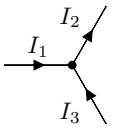
We can also keep track of the sense of "in" and "out" by keeping track of the *sign* of the current (always important to do!). Thus, for example, in the following junction,



we should write

$$I_1 + I_2 + I_3 = 0. \tag{1.9}$$

On the other hand, if we draw the currents like *this*,



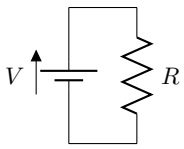
then we should write

$$I_1 - I_2 + I_3 = 0. \tag{1.10}$$

2. The **voltage law** states that around a *closed circuit* or *loop* in a circuit, the EMFs must balance the voltage drops, or

$$\sum_j \mathscr{E}_j = \sum_j V_j, \tag{1.11}$$

where the $\mathscr{E}_j$ are the EMFs, and the V_j are the voltage drops. For example, in the circuit below, the EMF is V due to the battery, so the voltage drop across the resistor must also be V.



1.3.1 Series Resistors

Two resistors in series are the same as a single resistor, with an effective resistance that is the sum of the individual resistors.

Why? (Try to work this out on your own!)

The idea is that any current I that flows through one must flow through the other. So the voltages across the resistors are $V_1 = IR_1$ and $V_2 = IR_2$. Then the total drop across both resistors is

$$V = V_1 + V_2 = I(R_1 + R_2) =: IR_{\text{eff}}. \tag{1.12}$$

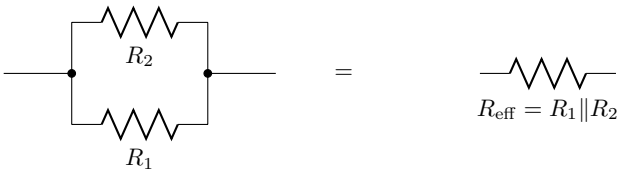
Thus,

$$R_{\text{eff}} = R_1 + R_2 \tag{1.13}$$

is the effective resistance of the series pair. Of course, this generalizes to multiple resistors.

1.3.2 Parallel Resistors

Two resistors in parallel also behave as a single resistor, as shown below.



The shorthand notation here is

$$\frac{1}{R_1 \| R_2} := \frac{1}{R_1} + \frac{1}{R_2}, \tag{1.14}$$

so that more parallelism in resistors *decreases* resistance.

Why is this? In this case, the *voltage* V is common to the two resistors. The currents are $I_1 = V/R_1$ and $I_2 = V/R_2$. But the total current through the pair must be $I = I_1 + I_2$, which satisfies $I = V/R_{\text{eff}}$. This means

$$\frac{V}{R_{\text{eff}}} = \frac{V}{R_1} + \frac{V}{R_2}, \tag{1.15}$$

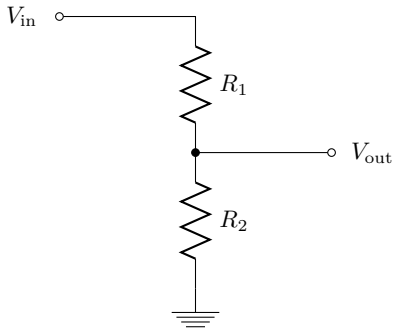
and canceling the voltage gives

$$\frac{1}{R_{\text{eff}}} = \frac{1}{R_1} + \frac{1}{R_2}, \tag{1.16}$$

which agrees with the shorthand above.

1.3.3 Voltage Divider

This is a useful combination of resistors that occurs *all the time* in circuits.



Here at the bottom of the circuit diagram, we are drawing the **ground** symbol, which means we are declaring this point to be a *fixed* voltage (say, zero), and all other voltages are *differences* with respect to ground. (In the "ground" or "earth" pin on an ac power receptacle, this is *literally* the ground outside. Often the case of electrical devices is connected to ground for safety, and in cars the entire chassis is ground.)

Now due to the input voltage V_{in}, some current flows in from the input. This must satisfy

$$I = \frac{V_{\text{in}}}{R_1 + R_2}. \tag{1.17}$$

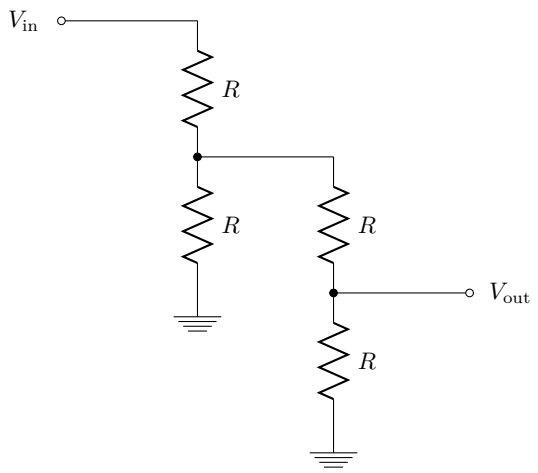
The same current flows through R_2, so

$$V_{\text{out}} = IR_2 = \left(\frac{R_2}{R_1 + R_2}\right) V_{\text{in}}. \tag{1.18}$$

(voltage divider)

The output voltage is reduced from the input by the ratio of R_2 to the total resistance. **This is important; you should memorize this.** Especially if this is made from an adjustable resistor (**potentiometer**), this can be used to make an adjustable voltage source.

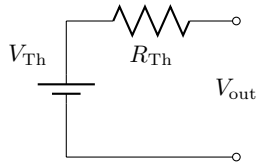
However, suppose we *chain* two voltage dividers, as follows.



Is the output voltage just divided twice, or 1/4 the input voltage? Why *not*?

1.4 Thévenin's Theorem

Thévenin's theorem is very useful in analyzing passive networks. It says that **any** network of resistors and EMFs—if we interact with it only at two points—can be replaced by an equivalent circuit of a series EMF and resistor, as shown.



The EMF and resistance are called the **Thévenin equivalent voltage** and **Thévenin equivalent resistance**, respectively. This equivalent circuit is a direct consequence of the *linearity* of a passive network.

How do we find the Thévenin equivalent component values? A couple of observations:

1. If nothing is connected to the output, then $V_{\text{out}} = V_{\text{Th}}$.
2. If the output is short-circuited,[1] then $V_{\text{out}} = 0$ and $V_{\text{Th}} = R_{\text{Th}} I_{\text{short}}$.

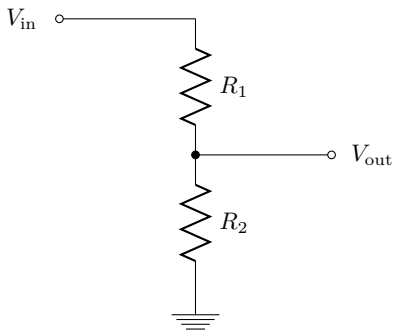
These allow you to infer the Thévenin values. The second rule is useful experimentally, provided shorting the output does not destroy the circuit!

An alternate way to find the Thévenin resistance, especially in analyzing a circuit on paper, is to replace all EMFs by short circuits, and compute the equivalent resistance at the output.

1.4.1 Voltage Divider

Back to the voltage divider of Section 1.3.3.

[1] A "short circuit" or "short" is a direct, low-resistance path between two points in a circuit, like a piece of wire. It acts like a "shortcut" for current between the two points, compared to going through the regular, higher-resistance paths in the circuit.



1. We already found the Thévenin voltage as the unloaded-output voltage:

$$V_{\text{Th}} = \left(\frac{R_2}{R_1 + R_2}\right) V_{\text{in}}. \tag{1.19}$$

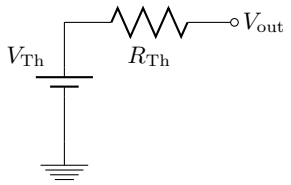
(voltage divider, Thévenin voltage)

2. If we short the output (to ground), the current only flows through R_1, so $I_{\text{short}} = V_{\text{in}}/R_1$. Then using $R_{\text{Th}} = V_{\text{Th}}/I_{\text{short}}$, we get

$$R_{\text{Th}} = \frac{R_1 R_2}{R_1 + R_2} = R_1 \| R_2.$$

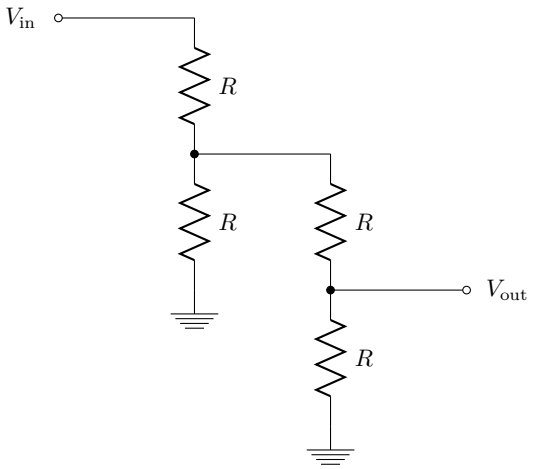
(voltage divider, Thévenin resistance) (1.20)

That is, the equivalent circuit is as below

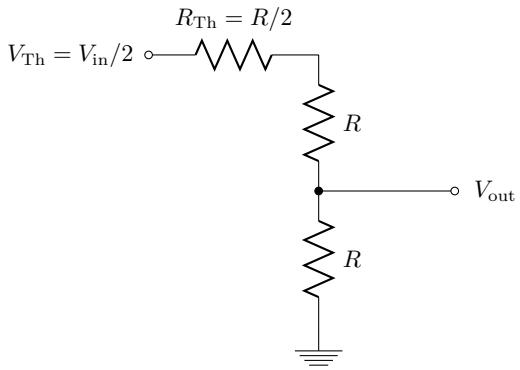


The equivalent voltage is the divided voltage, and the equivalent resistance is the **parallel resistance** of the two resistors. **This is important; you should memorize this.**

Now back to the example of two cascaded voltage dividers.



We can replace the first divider by the Thévenin equivalent:



Now we are back to a simple voltage divider, so

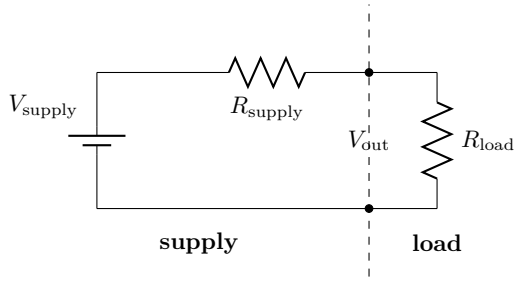
$$V_{\text{out}} = \left(\frac{2}{5}\right)\frac{V_{\text{in}}}{2} = \frac{V_{\text{in}}}{5}. \tag{1.21}$$

1.4.2 Connected Circuits and Power Transfer

The idea of Thévenin equivalence is also useful when analyzing what happens when you connect two fairly arbitrary circuits together. Suppose we take the common situation of one circuit "powering" another. That is, two circuits interact,

1. The "**supply**" is some circuit **with** EMFs.
2. The "**load**" is some circuit **without** EMFs.

Using the Thévenin-equivalent circuits for both, we can represent the connection thusly:



This is just a voltage divider, so we can write the output as

$$V_{\text{out}} = \left(\frac{R_{\text{load}}}{R_{\text{load}} + R_{\text{supply}}}\right) V_{\text{supply}}. \tag{1.22}$$

Now a few remarks are in order.

1. If $R_{\text{load}} \gg R_{\text{supply}}$, then $V_{\text{out}} \approx V_{\text{supply}}$, and the supply acts like an ideal voltage source. This applies to the "unloaded" (large R_{load}) and "good supply" (small R_{supply}) regimes.
2. If the supply is a battery, R_{supply} is the "internal resistance" of the battery. The internal resistance of a battery is larger for smaller or "used-up" batteries. The symptom is that the voltage sags when you try to draw current.
3. R_{supply} is called the **output impedance** of the supply. R_{load} is called the **input impedance** of the load.

4. The **impedance-matching condition** answer, under what conditions is maximum power transferred from source to load? The power in the load is

$$P_{\text{load}} = \frac{V_{\text{out}}^2}{R_{\text{load}}} = \frac{R_{\text{load}}}{(R_{\text{load}} + R_{\text{supply}})^2} V_{\text{supply}}^2. \tag{1.23}$$

Maximizing this via

$$\frac{d}{da}\left(\frac{a}{(a+b)^2}\right) = \frac{1}{(a+b)^2} - \frac{2a}{(a+b)^3} = 0, \tag{1.24}$$

which leads to $a = b$, we have the matching condition

$$R_{\text{load}} = R_{\text{supply}}. \tag{1.25}$$
(impedance-matching condition)

This is saying, for a *fixed* source impedance, the most power we can get out of the source and into the load is if the load impedance matches the supply impedance. In older tube amplifiers, this was an important consideration. For efficient matching to different speaker loads, amplifier output transformers would often have different "taps" for $4\,\Omega$, $8\,\Omega$, $16\,\Omega$, etc. speakers.

1.5 Matrix Solution of Resistor Networks

Have a look at the XKCD comic "Circuit Diagram,"[2] and enjoy (you'll recognize more stuff here as you learn more about electronics). Make sure to hover the cursor over the comic so you see the last joke.

Now look at part of the circuit labelled "Oh, so you think you're such a whiz at EE 201?" (If you can't access the circuit for whatever reason, it is a rat's next of resistors, and the idea is to find the equivalent resistance.) Randall Munroe was joking, but we'll develop a systematic way to handle this kind of problem, which you can use to tackle that mess without much difficulty.

1.5.1 Review of Linear Algebra

First, we're going to use some linear algebra (in practice, we will want the help of a computer), so let's review the notation. A **matrix** is a group of numbers indexed by two numbers. For example, we can write down a 2×2 matrix as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \tag{1.26}$$

We can refer to the whole matrix as $\mathbf{A}$. We can also refer an **element** (one of the entries) of the matrix as A_{ij}. Note that i refers to the element's *row*, while j refers to the element's *column*. We can write a **system of linear equations** as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \tag{1.27}$$

This is just another way to write down the pair of equations

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 &= b_1 \\ A_{21}x_1 + A_{22}x_2 &= b_2 \end{aligned} \tag{1.28}$$

and the shorthand notation for the matrix form is

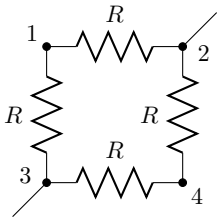
$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1.29}$$

where $\mathbf{x}$ and $\mathbf{b}$ are **vectors** (i.e., $n \times 1$ matrices, or specifically here, 2×1 matrices). Under certain conditions, it is possible to solve for the x_j in terms of the b_i and the A_{ij} (we'll let a computer help here). Make sure you understand the pattern of the matrix-vector multiplication in the equations above *before* you continue.

[2] `http://xkcd.com/730/`

1.5.2 Matrix Form of the Resistance Network and Example

To develop the method to solve general networks of resistors, we'll apply it to the simple circuit below, to make the method more intuitive.



The goal is to find the effective resistance between points 2 and 3.

Instead of working with resistance, we'll work with the **conductance** $G = 1/R$, so that Ohm's law is

$$I = GV. \tag{1.30}$$

Why? Well, conductance is well-defined between disconnected points ($G = 0$), and it will turn out to add up in the correct way for this formalism. Now between any two nodes in the network, we can write Ohm's law as

$$I_{ij} = G_{ij}(V_i - V_j), \tag{1.31}$$

where V_j is the voltage at node j, G_{ij} is the conductance of the connection between the two points (with $G_{jj} = 0$, so there is no "self-conductance", and $G_{ij} = G_{ji}$), and I_{ij} is the current flowing from node i to node j. Specifically, in this example problem, we have $G_{12} = G_{24} = G_{43} = G_{31} = G$, while $G_{23} = G_{14} = 0$.

Now Kirchhoff's current law states that the sum of currents flowing out of a node must be zero (note that we aren't yet considering any currents going *in* to a node, the way we defined I_{ij}):

$$\sum_{j \neq i} I_{ij} = 0. \tag{1.32}$$

Note that we explicitly excluded the $I_{ii} = 0$ case. This means

$$\sum_{j \neq i} G_{ij}(V_i - V_j) = 0. \tag{1.33}$$

Separating out terms we can rewrite this as

$$\left(\sum_{j \neq i} G_{ij} \right) V_i + \sum_{j \neq i} (-G_{ij}) V_j = 0. \tag{1.34}$$

We can interpret this as a linear system of equations as follows.

- The coefficient of V_i represents the diagonal elements of a matrix, call it $\mathbf{A}$. Then A_{ii} is the sum of all conductances connected to node i.
- The sum in the second term gives all the off-diagonal elements of the matrix. That is, $A_{ij} = -G_{ij}$, or the conductance between nodes i and j, with a minus sign.

To apply this to the example above, we will have a 4×4 matrix $\mathbf{A}$, and the system (1.34) of equations becomes

$$R^{-1} \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = 0, \tag{1.35}$$

or $\mathbf{Av} = 0$ for short. Note that R^{-1} multiplies every element in the matrix $\mathbf{A}$. Note also that the matrix A_{ij} is **symmetric**: $A_{ij} = A_{ji}$. **This is a very good sanity check for this matrix**, especially when you set up the matrix for part (c). Go through the elements of this matrix to make sure they make sense. The diagonal elements are all 2 because every node is connected to two identical resistors. There is a -1 representing every connection (via an identical resistor) between two nodes.

1.5.3 Solution for the Effective Resistance

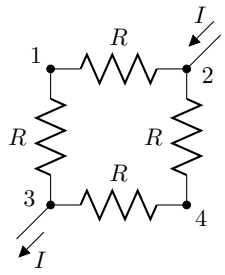
Unfortunately, the above matrix equation is not useful as is, because it is satisfied by $V_i = V$ for any constant V. Physically, without any applied voltage, this circuit really doesn't *do* much. But first, we'll take care of *another* problem: the matrix $\mathbf{A}$ is *singular*, meaning one of the equations in the linear system is redundant. Physically, this is because the absolute voltage of the circuit is not defined (the equations, by construction, only determine voltage *differences*). We can fix this by explicitly tying one of the nodes to zero voltage. Since node 3 is one of the nodes of interest, let's set

$$V_3 = 0, \tag{1.36}$$

which means we replace one of the equations in the linear system by this one. In matrix form, we will modify the third row of the matrix as follows:

$$R^{-1} \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = 0. \tag{1.37}$$

To solve the other problem (i.e., to make the circuit do something), let's introduce a current I, which flows into node 2. The same current I must flow *out* of node 3.



To handle this, we will modify Eq. (1.34) to say that the sum of currents flowing out of a node is equal to the currents flowing into a node (counting the external currents I as inputs, with minus signs to properly reflect their direction):

$$\left(\sum_{j \neq i} G_{ij} \right) V_i + \sum_{j \neq i} (-G_{ij}) V_j = I_{\text{in},i}. \tag{1.38}$$

Here $I_{\text{in},i}$ is the current flowing in to node i. In our example problem, $I_{\text{in},2} = I$, while $I_{\text{in},3} = -I$. Now writing the linear system including these currents, we have

$$\begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} 0 \\ IR \\ 0 \\ 0 \end{bmatrix}. \tag{1.39}$$

Note that we kept the third row of the matrix to reflect $V_3 = 0$, and we only introduced the current I at node 2. This is a well-defined system of equations, which we can solve to obtain V_2. The effective resistance of the network is defined by

$$R_{\text{eff}} = \frac{V_2 - V_3}{I} = \frac{V_2}{I}. \tag{1.40}$$

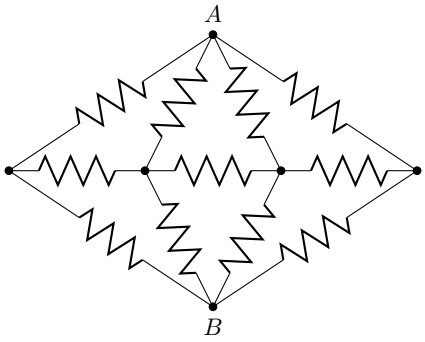
When you solve the matrix equation, as you might expect, the result should be $R_{\text{eff}} = R$. We will leave the details and the application to the XKCD problem to an exercise (Problem 17).

1.6 Circuit Practice

Here are a few example circuits to analyze with solutions; try to work these out and test your understanding so far. (Try *before* looking at the solutions!)

1.6.1 Reflection-Symmetric Network

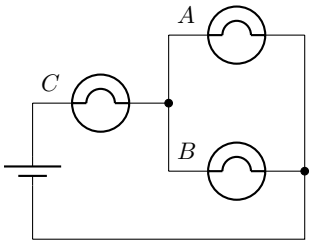
Given that all resistors here are equal and of resistance R, what is the equivalent resistance between A and B?



Solution. Notice that by symmetry, the voltage across the center resistance is zero, so the current flowing through it is zero, no matter what the voltage V_{AB}. Thus, we can remove it from the circuit. The same goes for the two other "equatorial" resistors. Thus, we have 4 parallel resistances of $2R$, for a total resistance $R/2$.

1.6.2 Series and Parallel Light Bulbs

Suppose the three bulbs in the circuit below are identical.



What is the relative brightness of the bulbs? Do bulbs B and C get brighter or dimmer when you remove A?

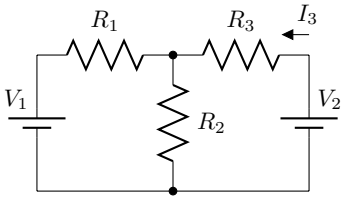
Solution. First, note that *half* the C current flows through A and B, while A and B drop half the voltage of C. So A and B are $1/4$ as bright as C.

When A is removed, then B and C are equally bright. Less current flows overall, because the total resistance is larger. C drops less voltage than it used to, B drops more than it used to. Clearly C should be dimmer, but B should be brighter, because it now has more voltage and more current (the drop in overall current is outweighed by the factor of 2 this bulb gets by its neighbor disappearing).

To be careful, note in the original case that the voltage across C is $(2/3)V$, and across A and B it is $(1/3)V$, where V is the battery voltage. The current through C is $(2/3)V/R$, and half that flows through A and B. In the new case, each bulb drops $V/2$, and current $(1/2)V/R$ flows through each. Before, the C power was the product of voltage and current, or $(4/9)V^2/R$, and now it is $(1/2)V^2/R$, so this is dimmer. Before, the B power was $(1/9)V^2/R$, so this one gets brighter.

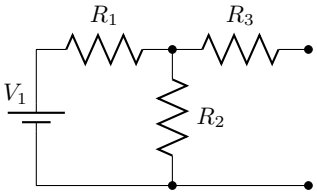
1.6.3 Thévenin Circuit

Compute I_3 in the circuit below, by using the Thévenin equivalent to the circuit, thinking of V_2 as the "load" for the rest of the circuit.



Use values $R_1 = 1\,\Omega$, $R_2 = 2\,\Omega$, $R_3 = 4\,\Omega$, $V_1 = 1\,\mathrm{V}$, and $V_2 = 2\,\mathrm{V}$.

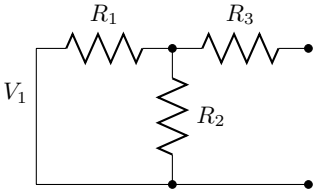
Solution. To work out the Thévenin equivalent, first look at the "unloaded" circuit:



The output voltage (and thus Thévenin-equivalent voltage) here is just the voltage-divider result

$$V_{\mathrm{Th}} = \left(\frac{R_2}{R_1 + R_2}\right) V_1 = \frac{2}{3}\,\mathrm{V}. \tag{1.41}$$

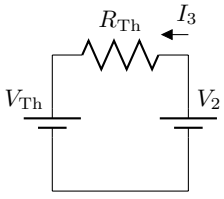
To get the resistance, replace the EMF by a short:



The Thévenin resistance is the resistance that appears at the output terminals, or

$$R_{\mathrm{Th}} = R_3 + R_1 \| R_2 = R_3 + \frac{R_1 R_2}{R_1 + R_2} = \frac{14}{3}\,\mathrm{V}. \tag{1.42}$$

Now going back to the original circuit, we have the equivalent



This is pretty easy to solve, just use Ohm's law:

$$I_3 = \frac{V_2 - V_{\mathrm{Th}}}{R_{\mathrm{Th}}} = \frac{(4/3)\,\mathrm{V}}{(14/3)\,\Omega} = \frac{2}{7}\,\mathrm{A}. \tag{1.43}$$

1.7 Exercises

Problem 1.1

Ohm's law for the **conductance** G is

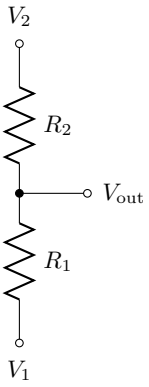
$$I = GV, \tag{1.44}$$

where the conductance is related to the resistance by $G = 1/R$, and G is measured in *mhos*, or ℧. (An equivalent but less entertaining unit is the *siemens*, or S.)

Derive expressions for the conductance G of two conductors of conductances G_1 and G_2 in series. Do the same for two conductors in parallel. Use only the form of Ohm's law above to start; do *not* use the analogous results for parallel and series resistances.

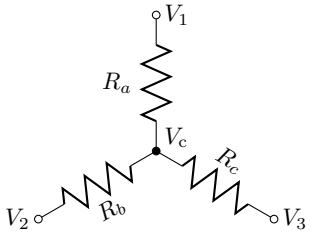
Problem 1.2

Compute V_{out}. This is a voltage divider, but the "bottom" end is not grounded. *Put the result into some form you can remember and then memorize it.*



Problem 1.3

In this circuit,

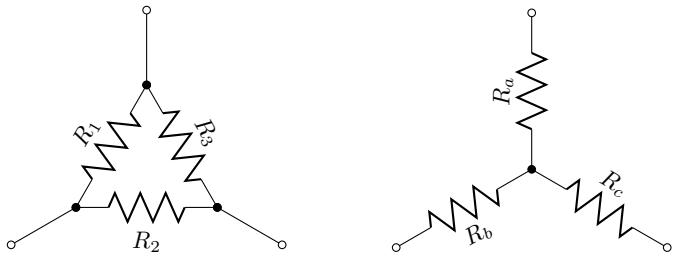


show that the "center voltage" V_c is given by

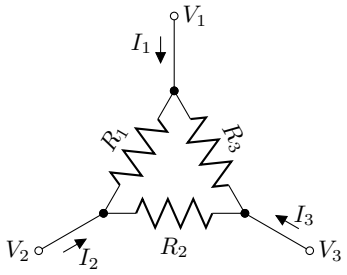
$$V_c = \frac{R_b R_c V_1 + R_c R_a V_2 + R_a R_b V_3}{R_a R_b + R_b R_c + R_c R_a}. \tag{1.45}$$

Problem 1.4

Find relations between resistances R_1, R_2, and R_3 in the left-hand network, and resistances R_a, R_b, and R_c in the right-hand network, that make the two networks equivalent.



What does it mean for two resistor networks to be equivalent? Label the voltages and currents, for example, as follows.



Now it's useful to think of some of the voltages and currents as "inputs," and others as "outputs," or "responses." For example, if the three voltages are inputs, then they cause the corresponding currents to flow. You can also think of one or two of the currents as being inputs instead of their respective voltages. (But *not* all three currents; why?) The circuits are equivalent if every set of inputs produces the equivalent set of outputs. For the purposes of this problem, it is sufficient to show that the two networks produce equivalent relations between currents and voltages.

To do this:

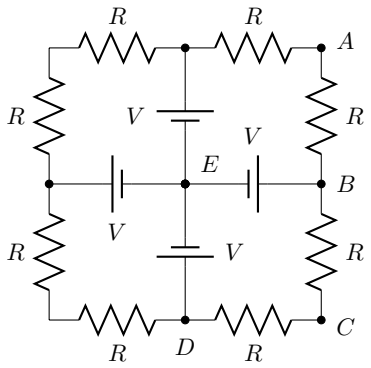
(a) Derive expressions for the currents I_1, I_2, and I_3 in terms of the voltages for the left-hand ("Delta") circuit.

(b) Derive expressions for the currents I_1, I_2, and I_3 in terms of the voltages for the right-hand ("star") circuit.

(c) Set the corresponding currents in the two circuits equal to each other, and derive expressions for R_a, R_b, and R_c in terms of R_1, R_2, and R_3. (You may think of the voltages as "inputs," and thus as independent parameters.)

Problem 1.5

Consider the circuit below.



Give the Thévenin-equivalent circuit for the circuit in the diagram, as seen between

(a) points B–D.

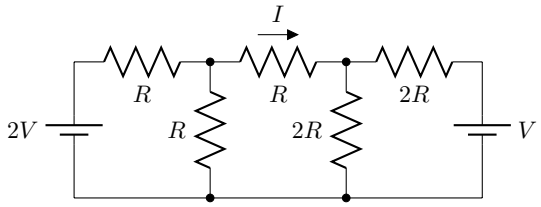
(b) points A–B.

(c) points B–E.

(d) points A–E.

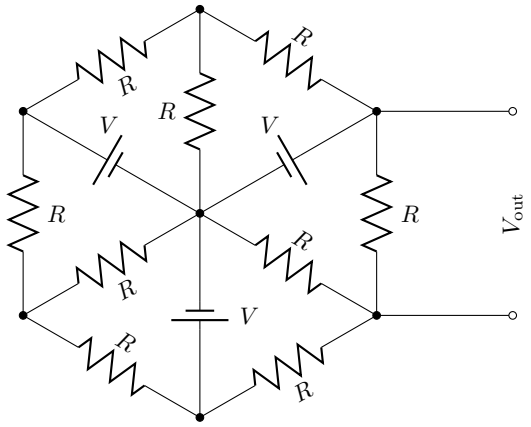
Problem 1.6

In the circuit below, compute the current I through the top-middle resistor.



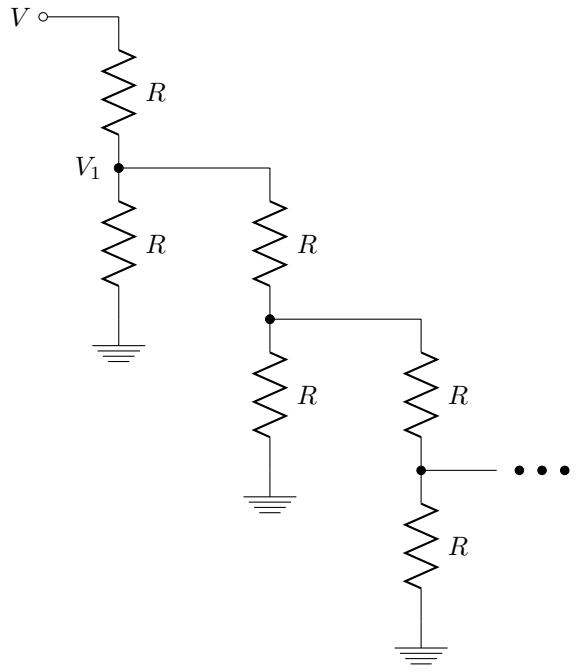
Problem 1.7

Give the Thévenin voltage V_{Th} and resistance R_{Th} for the Thévenin equivalent circuit of the circuit shown below.

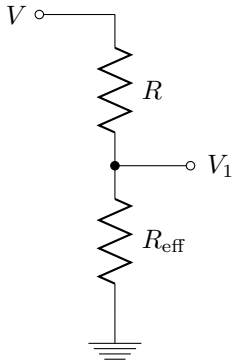


Problem 1.8

Consider the following circuit, consisting of an infinite cascade of voltage dividers.



By how much does the rest of the divider chain load down the first divider? That is, in the equivalent circuit

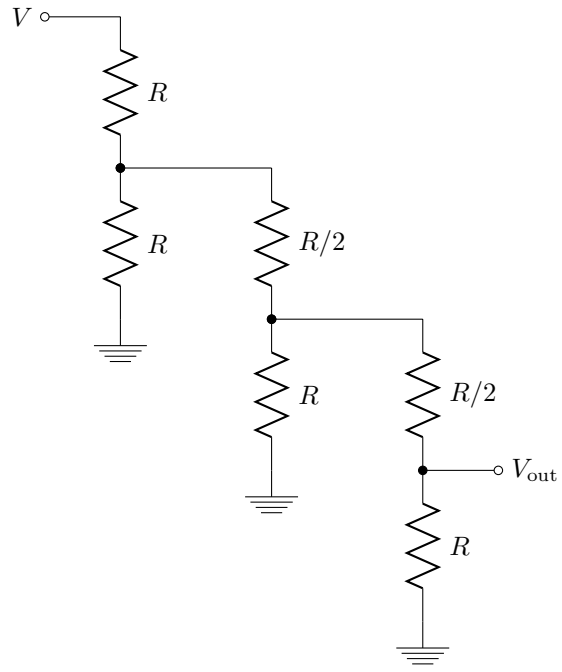


what is the value of R_{eff}?

Hint: In the original divider chain, consider the *second* divider, along with the rest of the chain. *This* part of the circuit is *also* equivalent to the same effective circuit. Now you have two different, equivalent circuits in terms of R and R_{eff}; compute the resistance from the point V to ground in both circuits, equate, and solve for R_{eff}.

Problem 1.9

Consider the circuit below, with 3 cascaded voltage dividers (not all the same).

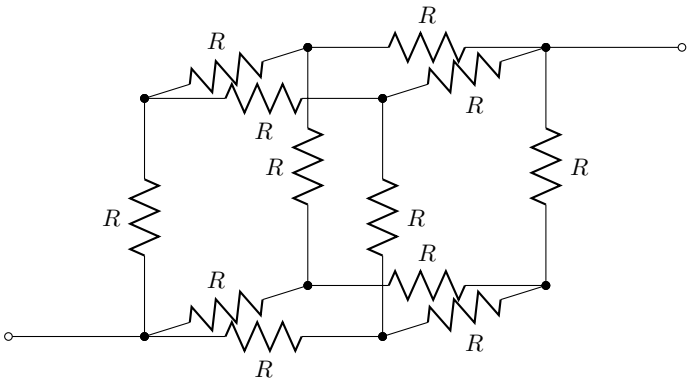


(a) Compute V_{out}.

(b) Compute the current in *each one* of the resistors in the circuit, assuming no load connected to V_{out}.

Problem 1.10

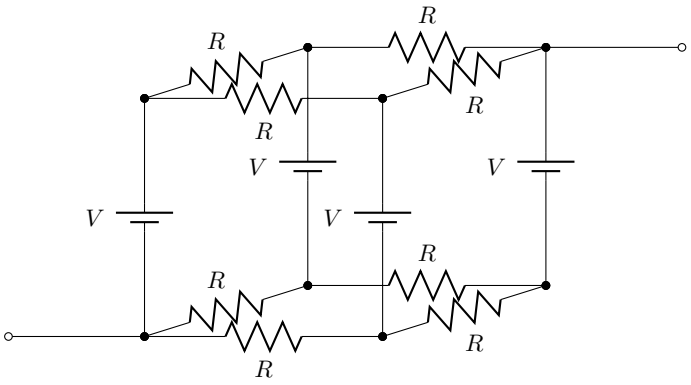
Consider a network of 12 identical resistors of resistance R in the shape of a cube, as shown in the circuit below. Compute the equivalent resistance between the two terminals shown at opposite corners of the cube.



Hint: use symmetry, and analyze a current flowing between the two terminals. It also may (or may not) help to identify which junctions are at equivalent potentials.

Problem 1.11

Consider the network below of 8 identical resistors and 4 identical voltage sources. With the two indicated terminals as the output, give the Thévenin equivalent circuit. *Explain* your result.



Problem 1.12

A "schmesistor" is a device that obeys "Schmohm's law,"

$$V = I^2 S, \tag{1.46}$$

where S is the "schmesistance."

(a) Show that two schmesistors S_1 and S_2 in **series** behave like one schmesistor with schmesistance $S_{\text{eff}} = S_1 + S_2$.

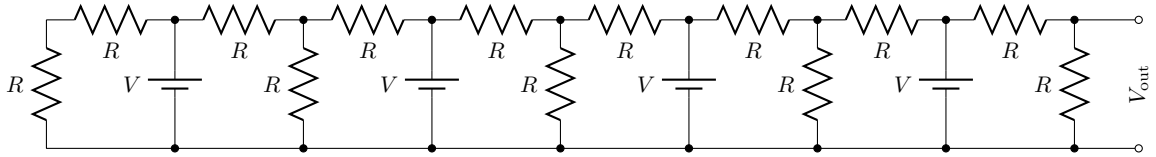
(b) Show that two schmesistors S_1 and S_2 in **parallel** behave like one schmesistor with schmesistance

$$S_{\text{eff}} = \left(\frac{1}{\sqrt{S_1}} + \frac{1}{\sqrt{S_2}} \right)^{-2}. \tag{1.47}$$

(c) Describe the difficulty in using complex notation to analyze a circuit that includes schmesistors.

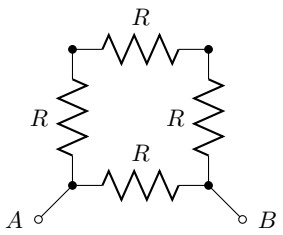
Problem 1.13

Give the Thévenin equivalent circuit for the circuit below, as "seen" by the output terminals marked by V_{out}.

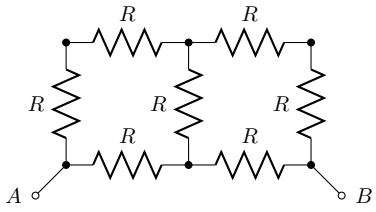


Problem 1.14

(a) Compute the effective resistance between points A and B.

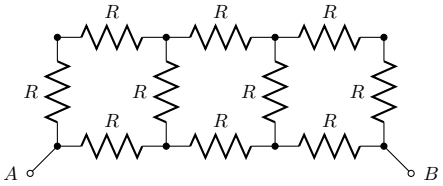


(b) Compute the effective resistance between points A and B.



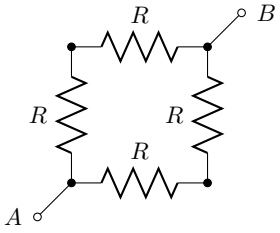
Hint: **symmetry**.

(c) Compute the effective resistance between points A and B.

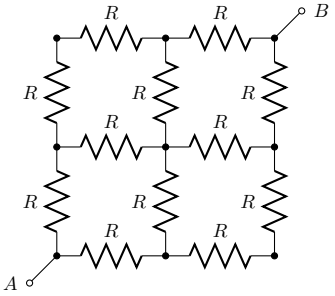


Problem 1.15

(a) Compute the effective resistance between points A and B.

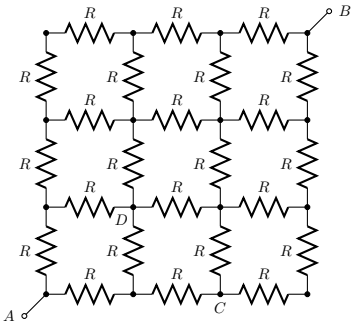


(b) Compute the effective resistance between points A and B.



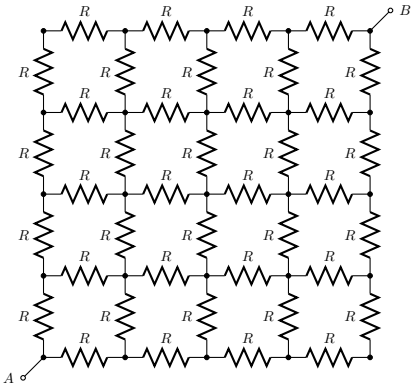
Hint: **symmetry**. There are (at least) two symmetries you should be taking advantage of here.

(c) Compute the effective resistance between points A and B.



Hint: **symmetry**, but note that it is *not* true in general that $V_C = V_D$.

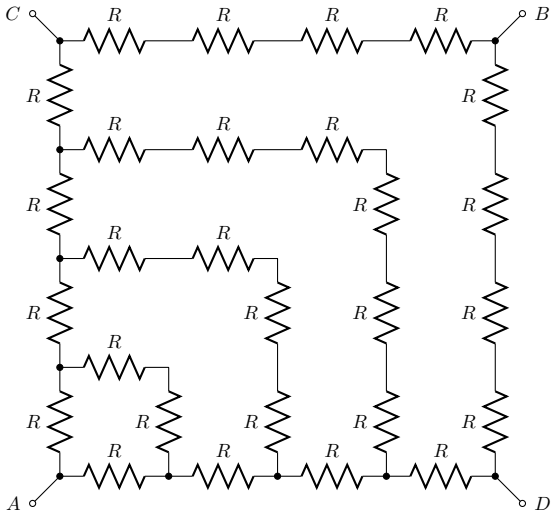
(d) Generalize (a)–(c) to a resistor network forming a 4×4 grid(!) of squares.



Problem 1.16

(a) Compute the equivalent resistance between points A and B in the circuit below.

(b) Compute the equivalent resistance between points C and D in the circuit below.



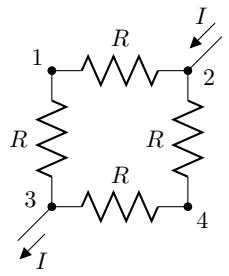
Problem 1.17

Now back to the XKCD comic "Circuit Diagram."[3] The point of this problem is to apply the matrix method of Section 1.5 to solve this problem.

(a) Solve the example linear system in Eq. (1.39), using a computer, to obtain a value for V_2, and thus for R_{eff}. I will describe how you can do this in *Mathematica* as an example (which is obnoxiously expensive but handy in that it will give *exact* results using this method), but a similar procedure will work in some other software packages as well, such as the open-source *Octave*.[4]

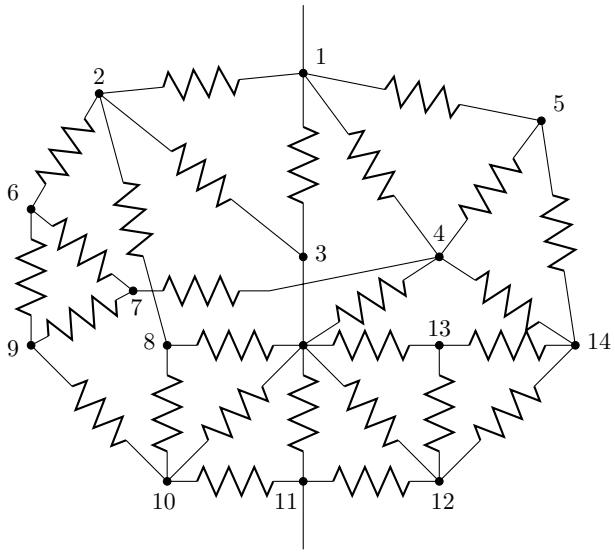
First in a *Mathematica* window, type "`A=`", and then under the menu `Insert > Table/Matrix > New...`, create a 4×4 matrix. Then fill in the elements as above. Do the same thing to create a vector (4×1 matrix) for the currents (call it "`II`", and use "`IR`" to represent IR in the vector; *Mathematica* reserves the symbol "`I`" for $\sqrt{-1}$). Finally, use `LinearSolve[A, II]` to obtain the solution vector $\mathbf{v}$, and use it to deduce the effective resistance.

(b) Modify the calculation to compute the effective resistance between nodes 3 and 4 in the same example circuit (reproduced below).



(c) Finally, modify the calculation to handle the XKCD mess.

The network is reproduced below, with nodes labeled (all 1-Ω resistors). You should set $V_{11} = 0$ and let a current I flow into node 1 (and out node 11).



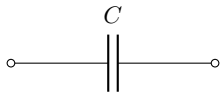
[3] `http://xkcd.com/730/`
[4] `http://octave.org`

Chapter 2

Capacitors and Inductors

2.1 Capacitor Basics

A **capacitor** (an older, equivalent term is **condensor**) is, at minimum, a pair of conductors separated by vacuum or dielectric. The symbol itself depicts two parallel electrodes.



Generally, useful capacitors have a relatively large area to achieve a reasonably useful capacitance, so they are something like a pair of planar conductors, with little separation. They may have stacks of many planar conductors to increase the area even further, and the stacks may be rolled up to save space and stored in a can or dipped in epoxy for robustness.

Capacitors act as devices to store *charge*, and in doing so they also store *energy*. The charge stored on the plates generates an electric field and thus a potential difference between the capacitor plates. The potential and charge are related by the capacitor law,

$$Q = CV, \tag{2.1}$$
(capacitor law)

where C is the **capacitance**, measured in **Farads (F)**, which characterizes the capacitor. Larger conductor areas, "stronger" dielectrics, and smaller electrode spacing all result in larger capacitance. (For example, a few twists to intertwine two pieces of insulated hookup wire makes a capacitor of a few pF.) Differentiating this relation at constant capacitance, and using $I = dQ/dt$, we find

$$I = C\frac{dV}{dt}, \tag{2.2}$$
(capacitor charging)

which means that a current *charges* or *discharges* a capacitor, allowing it to build up or bleed off charge, thus changing the voltage.

There are many types of capacitors, and usually the different types are named according to their dielectrics. A few of the most important are:

1. **ceramic/monolithic**: these tend to be cheap, and work reasonably well at high frequencies. The capacitances are fairly small in the overall spectrum, ranging from a $\sim$pF to about $\sim 0.1\,\mu$F.

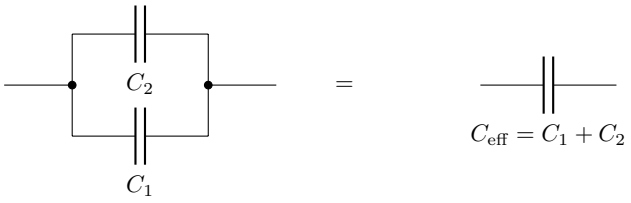
2. **electrolytic**: generally this refers to **aluminum electrolytic capacitors**. In these capacitors, one electrode is aluminum foil, and the other is a liquid electrolyte. Under normal operation, an oxide layer grows on the aluminum and acts as a thin dielectric layer. The foil can be coiled to give large surface area, and the dielectric is very thin, so capacitances can be large, from $\sim 1\,\mu$F to $\sim 1\,$F or more,

though these tend not to work as well at high frequencies. These are **polarized**, meaning they can only sustain voltage applied in one direction; the wrong voltage polarity can cause the insulating layer to break down, leading to failure of the capacitor.

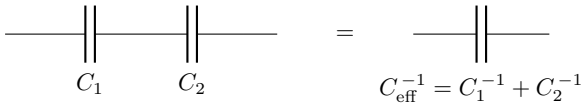
3. **tantalum**: these are also electrolytic capacitors, with a sintered titanium pellet as one electrode, and a solid electrolyte as the other. These typically have intermediate capacitances in the range of $\sim 0.1\,\mu$F to $\sim 10\,\mu$F, and are fairly compact compared to aluminum electrolytics.

There are many other kinds of capacitors. Some examples include paper, mica, mylar, polystyrene, polypropylene, oil, and niobium electrolytic.

Like resistors, networked capacitors can combine to form equivalent single capacitors. Two capacitors in *parallel* simply add their capacitances,



as resistors in series add. Two *series* capacitors add less straightforwardly,

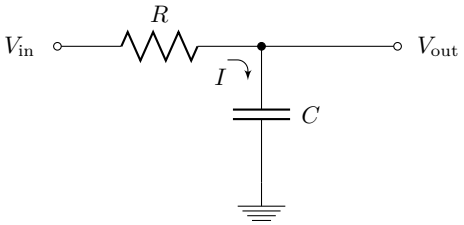


like two *parallel* resistors add.

2.2 Simple R–C Circuits

2.2.1 Integrator

Consider the simple R–C circuit below. It looks something like a voltage divider, with the lower resistor replaced by a capacitor.



To analyze this (unloaded) circuit, note that all of the current I going through the resistor must also pass through the capacitor. Applying Ohm's law to determine the current,

$$V_{\text{in}} - V_{\text{out}} = IR, \tag{2.3}$$

and then using the capacitor-charging law (2.2) to relate the current to the output voltage,

$$I = C\frac{dV_{\text{out}}}{dt}. \tag{2.4}$$

Solving for dV_{out}/dt and eliminating I,

$$\frac{dV_{\text{out}}(t)}{dt} = -\frac{V_{\text{out}}(t)}{RC} + \frac{V_{\text{in}}(t)}{RC}. \tag{2.5}$$

(differential equation for integrator)

Note that by comparing left- and right-hand sides here, we can see that RC must have the units of time for the units to work out.

2.2.1.1 Solution by Integrating Factor

So how do we solve this differential equation? Note that the input $V_{\text{in}}(t)$ is arbitrary and unknown, so the best we can do is to obtain the solution in terms of $V_{\text{in}}(t)$. For this type of **ordinary differential equation (ODE)**, there is a nice trick to simplify it into something more manageable. The idea is to first set $V_{\text{in}} = 0$, and look at the equation.

$$\frac{dV_{\text{out}}(t)}{dt} = -\frac{V_{\text{out}}(t)}{RC}. \tag{2.6}$$

This ODE is easy to solve: it's just exponential decay,

$$V_{\text{out}}(t) = V_{\text{out}}(0)\, e^{-t/RC}. \tag{2.7}$$

Now the idea is that even in the *presence* of V_{in}, we should expect more or less the *same* (exponential-decay) behavior. So let's "build this in" to the solution by assuming a solution of the form

$$V_{\text{out}}(t) = \tilde{V}(t)\, e^{-t/RC}, \tag{2.8}$$

where $\tilde{V}(t)$ is the "deviation" from the simple solution $e^{-t/RC}$, which is called an **integrating factor**. (We aren't losing anything in this assumption, because $\tilde{V}$ could be anything.) Then solving for $\tilde{V}$,

$$\tilde{V} = V_{\text{out}}\, e^{t/RC}, \tag{2.9}$$

and differentiating,

$$\frac{d\tilde{V}}{dt} = \frac{dV_{\text{out}}}{dt}\, e^{t/RC} + \frac{1}{RC} V_{\text{out}}\, e^{t/RC}. \tag{2.10}$$

Now if we multiply Eq. (2.5) by $e^{t/RC}$, and bring both V_{out} terms to the left,

$$\frac{dV_{\text{out}}}{dt}\, e^{t/RC} + \frac{V_{\text{out}}}{RC}\, e^{t/RC} = \frac{V_{\text{in}}}{RC}\, e^{t/RC}, \tag{2.11}$$

then notice the left-hand side is the right-hand side of Eq. (2.10). Thus, we have

$$\frac{d\tilde{V}}{dt} = \frac{V_{\text{in}}}{RC}\, e^{t/RC}. \tag{2.12}$$

Integrating both sides from 0 to t,

$$\tilde{V}(t) - \tilde{V}(0) = \frac{1}{RC} \int_0^t V_{\text{in}}(t')\, e^{t'/RC}\, dt'. \tag{2.13}$$

Using Eq. (2.9) to get rid of $\tilde{V}$, and multiplying through by $e^{-t/RC}$,

$$V_{\text{out}}(t) = V_{\text{out}}(0)\, e^{-t/RC} + \frac{1}{RC} \int_0^t V_{\text{in}}(t')\, e^{(t'-t)/RC}\, dt'. \tag{2.14}$$
(integrator solution)

Thus, we have a solution as an integral to Eq. (2.5) in terms of an arbitrary input voltage.

2.2.1.2 Constant Input: Exponential Charging

Now let's take the simple case where the capacitor is initially uncharged [$V_{\text{out}}(0) = 0$], and some **constant** voltage V_{in} appears at the input. Then it comes out of the integral, and we have

$$\begin{aligned} V_{\text{out}}(t) &= \frac{V_{\text{in}}}{RC}\, e^{-t/RC} \int_0^t e^{t'/RC}\, dt' \\ &= V_{\text{in}}\, e^{-t/RC}\, e^{t'/RC}\Big|_0^t, \end{aligned} \tag{2.15}$$

with final solution

$$V_{\text{out}}(t) = V_{\text{in}}\left(1 - e^{-t/RC}\right). \tag{2.16}$$
(integrator solution, constant input)

This is an exponential rise from 0 to V_{in}, with **time constant** RC. That is, looking at $e^{-t/RC}$ when $t = RC$, this falls from unity to $1/e$, or about 37%. So $1 - e^{-t/RC}$, when $t = RC$, rises from 0 to $1 - 1/e$, or about 63%.

One important thing to note about this is that the capacitor tends to **smooth** the input. Here, we can regard the problem as an input of a sudden voltage step at $t = 0$ [which is consistent with $V_{\text{out}}(0) = 0$], and the output is now smoothed over a time scale RC due to the exponential action of the R–C circuit.

2.2.1.3 Integration

This R–C circuit is called a **passive integrator** or **integrator**. The reason is as follows. Going back to Eq. (2.5), if $V_{\text{out}} \ll V_{\text{in}}$, then we can ignore the V_{out} term:

$$\frac{dV_{\text{out}}(t)}{dt} \approx \frac{V_{\text{in}}(t)}{RC}. \tag{2.17}$$

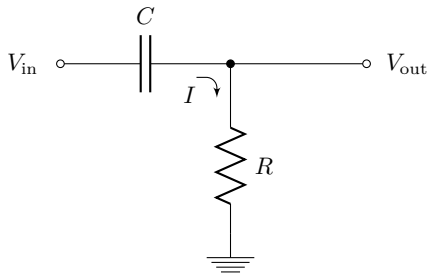
In this case, we can just integrate this equation directly from 0 to t:

$$V_{\text{out}}(t) - V_{\text{out}}(0) \approx \frac{1}{RC}\int_0^t V_{\text{in}}(t')\,dt'. \tag{2.18}$$

Thus, the output is the simple integral of the input signal, up to an offset and a factor of $1/RC$. In this regime, the capacitor simply stores up the charge that comes in from the resistor. If the capacitor charge, and hence voltage, becomes *too* large, then this reduces the voltage drop across the resistor, and the simple integral approximation breaks down.

2.2.2 Differentiator

We can get the other simple R–C circuit by interchanging the resistor and capacitor in the integrator, as shown below.



The capacitor charging equation here gives

$$I = C\frac{d}{dt}\big(V_{\text{in}} - V_{\text{out}}\big), \tag{2.19}$$

while Ohm's law gives

$$I = \frac{V_{\text{out}}}{R}. \tag{2.20}$$

Eliminating I, we get

$$\frac{V_{\text{out}}}{R} = C\frac{d}{dt}\big(V_{\text{in}} - V_{\text{out}}\big). \tag{2.21}$$
(differentiator ODE)

which is the ODE for this circuit. This circuit is called a **differentiator**, for reasons analogous to what we saw for the integrator. If $V_{\text{out}} \ll V_{\text{in}}$, then the ODE reduces to

$$\frac{V_{\text{out}}}{R} \approx C\frac{dV_{\text{in}}}{dt}, \tag{2.22}$$

and the output is approximately the derivative of the input.

It is possible to write down a general solution to Eq. (2.21) for the differentiator output in terms of an arbitrary input:

$$V_{\text{out}}(t) = V_{\text{in}}(t) + \big[V_{\text{out}}(0) - V_{\text{in}}(0)\big]e^{-t/RC} - \frac{e^{-t/RC}}{RC}\int_0^t V_{\text{in}}(t')\, e^{t'/RC}\, dt'. \tag{2.23}$$

(differentiator solution)

The idea is fairly similar to the integrator (the same integrating-factor trick applies); we will leave this as an exercise.

2.3 AC Signals and Complex Notation

Suppose we have an **ac signal** (alternating-current signal), oscillating at a single frequency ω, described by voltage:

$$V(t) = V_0 \cos \omega t. \tag{2.24}$$

Remember to distinguish *angular frequencies* from "regular frequencies." Angular frequencies are usually represented by ω, and are measured in rad/s. "Regular" frequencies are usually represented by f or ν, and are measured in Hz or cycles/s. We can write

$$\omega = 2\pi f \tag{2.25}$$

(angular frequency)

to relate the two. In physics, using the angular frequency saves writing a bunch of factors of 2π, but when quoting a physical value, it's best to stick to regular frequencies. That, is you could quote an angular frequency by saying "$\omega/2\pi = 100\,\text{Hz}$" instead of quoting the direct value of ω in rad/s.

As an example, let's go back to the capacitor charging law,

$$I = C\frac{dV}{dt}. \tag{2.26}$$

Then with the above ac signal, we have

$$I(t) = -\omega C V_0 \sin \omega t = \omega C V_0 \cos(\omega t + \pi/2), \tag{2.27}$$

where in the last step, we changed the negative sine to a phase-shifted cosine. This makes the current easier to compare to the original voltage. In fact, we can see that the current **leads** the voltage in phase, because the current's phase is larger by $\pi/2$ than the voltage's phase. The *shape* of the current signal is otherwise the same as the voltage, except for amplitude and phase.

2.3.1 Complex Phase

There is a nicer way to handle this monochromatic time dependence, and that is to introduce a complex-number notation. The idea is to represent the time dependence at frequency ω by a factor $e^{-i\omega t}$.[1] Then define a **complex voltage**

$$\tilde{V} = V_0 e^{-i\omega t}. \tag{2.28}$$

[1] Note that this is a common convention in *physics*; engineers usually use $e^{j\omega t}$, where $j = -i$. The physics notation comes from the time dependence of a right-going plane wave, $e^{i(kx-\omega t)}$.

The *real* (physical) voltage is just the real part of this, or

$$V(t) = \mathrm{Re}\big[\tilde{V}\big] = \mathrm{Re}\big[V_0 e^{-i\omega t}\big] = V_0 \cos\omega t, \tag{2.29}$$

if V_0 is real. The imaginary part, $iV_0 \sin\omega t$, is "carried along" for mathematical convenience, and should be dropped at the end of the calculation to get physical results. Thus, this works for *linear* circuits (in a *nonlinear* circuit, a single frequency would be converted into other frequencies, so this analysis is best for a capacitor or resistor, but less so for a diode).

Why is this representation convenient? First, *other* phases are easy to represent as the phase of the *complex* coefficient V_0. For example, we can write a phase ϕ as

$$\tilde{V} = V_0 e^{-i\omega t - i\phi}, \tag{2.30}$$

and the real part of this is

$$V(t) = V_0 \cos(\omega t + \phi), \tag{2.31}$$

as we expect for this phase. However, if we absorb the phase ϕ into the *complex* voltage amplitude V_0, then $\tilde{V}$ just looks like $V_0 e^{-i\omega t}$.

The other nice thing about this complex notation is that derivatives, integrals, and their associated phases are very easy to handle. We are assuming all time dependence is of the form $e^{-i\omega t}$. Then a time derivative acts on the phase of any object like

$$\frac{d}{dt} e^{-i\omega t - i\phi} =, -i\omega e^{-i\omega t - i\phi}. \tag{2.32}$$

Since this always happens, we can formally identify

$$\frac{d}{dt} \equiv -i\omega, \tag{2.33}$$
(derivative for monochromatic signals)

provided we are discussing monochromatic signals. Then the capacitor-charging rule (2.26) becomes

$$\tilde{I} = -i\omega C \tilde{V} \tag{2.34}$$

in complex notation.

2.3.2 Capacitive Reactance

We can interpret the complex form of the capacitor-charging rule in a powerful way. Solving for $\tilde{V}$, we find

$$\tilde{V} = \tilde{I} \frac{i}{\omega C}. \tag{2.35}$$

In fact, this looks a lot like Ohm's law, if we lump everything next to the $\tilde{I}$ into an effective "resistance,"

$$X_C := \frac{i}{\omega C}. \tag{2.36}$$
(capacitive reactance)

This is called the **capacitive reactance**, and has the same units as resistance. It functions as something like a resistance in the "Ohm's law for capacitors,"

$$\tilde{V} = \tilde{I} X_C, \tag{2.37}$$
(Ohm's law for capacitors)

which is the same as Eqs. (2.35) and (2.26). However, because the reactance represents a derivative, it depends on frequency. In fact, what this is saying is that capacitors have very high "resistance" at small frequencies (a capacitor is basically a broken wire, after all, and no current flows once the capacitor is charged), but the capacitor acts like a short circuit at high frequencies, as we will see.

2.3.3 Inductive Reactance

The same idea applies to inductors, which satisfy the "inductive-kick law"

$$V = L\frac{dI}{dt}, \tag{2.38}$$

in terms of the **inductance** L. Switching to complex notation and replacing the derivative, we have

$$\tilde{V} = -i\omega L\tilde{I}, \tag{2.39}$$

in which case we can define an **inductive reactance**

$$X_L := -i\omega L, \tag{2.40}$$
(inductive reactance)

so the inductor law becomes

$$\tilde{V} = X_L\tilde{I}, \tag{2.41}$$
(inductive reactance)

or just Ohm's law with X_L standing in for a resistance.

2.3.4 Impedance

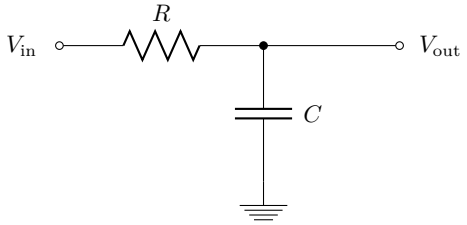
Since reactances and resistances "look" the same once they're stuck into Ohm's law, we can use the more general notion of **impedance** to represent any of these. To summarize the important points,

- **Resistances** R are always *real*.
- **Reactances** X_C, X_L are always *purely imaginary*, and they depend on frequency as well.
- An **impedance** can be any combination of resistances and reactances, and can be any complex value, not necessarily purely real or purely imaginary.
- Resistances and reactances are *special cases* of impedances.

The point of all this: for capacitors, inductors, and resistors in ac circuits, *everything* we did for resistive networks carries over to the ac case, in terms of impedances. This includes all the parallel, series, and Thévenin stuff.

2.3.5 Low-Pass Filter

As an example, let's return to the integrator circuit from Section 2.2.1.



If we think of the capacitor as being a resistor with "resistance" X_C, then this is just a voltage divider. Using the voltage-divider formula (1.18),

$$\tilde{V}_{\text{out}} = \frac{X_C}{R + X_C}\tilde{V}_{\text{in}} = \frac{1}{1 - i\omega RC}\tilde{V}_{\text{in}}. \tag{2.42}$$

This is a linear relation between the input and output voltage amplitudes, so to simplify the discussion a bit, let's define the **transfer function**

$$\tilde{T}(\omega) := \frac{\tilde{V}_{\text{out}}}{\tilde{V}_{\text{in}}}, \tag{2.43}$$
(transfer function)

which for the low-pass filter is

$$\tilde{T}(\omega) = \frac{1}{1 - i\omega RC} \tag{2.44}$$
(transfer function)

from Eq. (2.42). To simplify even more, we can also consider the **amplitude transfer function**, which discards the phase information:

$$T(\omega) := \left| \frac{\tilde{V}_{\text{out}}}{\tilde{V}_{\text{in}}} \right|. \tag{2.45}$$
(amplitude transfer function)

Then for the low-pass filter, from Eq. (2.42) we have

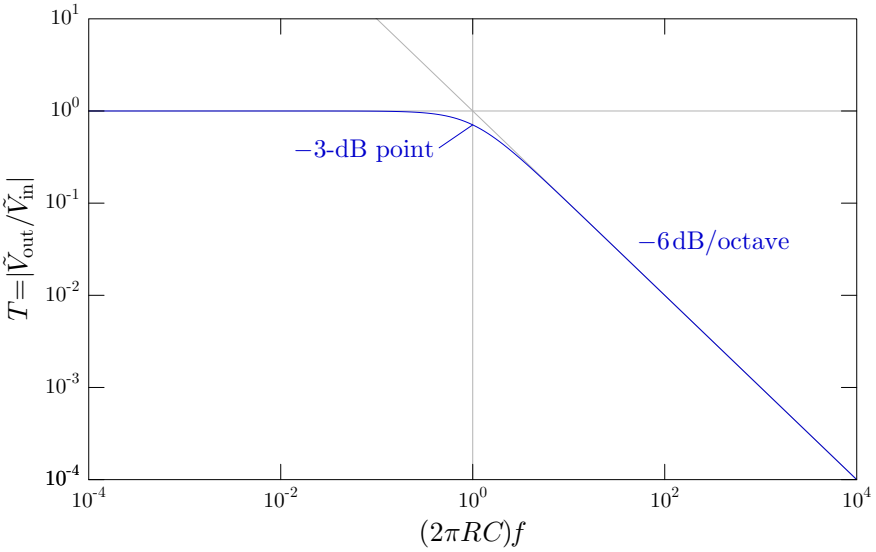
$$T(\omega) = \frac{1}{\sqrt{1 + (\omega RC)^2}}.$$
(amplitude transfer function, low-pass filter) (2.46)

Looking at the asymptotics of the low-pass filter, note that

- As $\omega \longrightarrow 0$ ($\omega \ll 1/RC$), note that $\tilde{T}(\omega) \longrightarrow 1$, which means that the integrator does not change the signal at low frequencies.
- For large ω ($\omega \gg 1/RC$), $\tilde{T}(\omega) \sim i/\omega RC$. Note that $i = e^{i\pi/2}$, and so $ie^{-i\omega t} = e^{-i(\omega t - \pi/2)}$. This means the output phase **lags** the input phase by 90°. Also, the output amplitude is reduced by a factor ω^{-1}. This **power-law behavior** appears as a straight line on a log–log plot, with slope -1.
- The high-frequency scaling of ω^{-1} is usually called $\mathbf{-6\,dB/octave}$. Remember that decibels are defined such that the ratio of two powers in decibels is of the form $10 \log_{10}(P/P_0)$, and the ratio of two *amplitudes* is $20 \log_{10}(V/V_0)$. One **octave** means a doubling of frequency (from the musical term), and ω^{-1} scaling means that doubling the frequency cuts the amplitude in half. In dB, this is $20 \log_{10}(1/2) = -6\,\text{dB}$.
- The **transition point** between the low- and high-frequency behavior is called the **3-dB point**, or more properly, the **−3-dB point**. The convention is to define the transition point as the point where $T(\omega)$ drops from 1 to $1/\sqrt{2}$ (so that the *power* transferred drops to 1/2). Then $20 \log(1/\sqrt{2}) = -3\,\text{dB}$. We can find the corresponding frequency by setting $T(\omega_{3\,\text{dB}}) = 1/\sqrt{2}$, which has the solution

  $$\omega_{3\,\text{dB}} = \frac{1}{RC}, \qquad f_{3\,\text{dB}} = \frac{1}{2\pi RC}. \tag{2.47}$$
  (3-dB frequency)

  All this is summarized in the plot below.



Because the integrator "passes" low frequencies without attenuation, and "rolls off" high frequencies, it is called a **low-pass filter**.

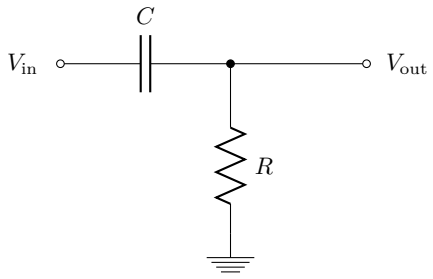
2.3.6 Example Problem: Alternate Scaling

What is the scaling of $-6\,\text{dB/octave}$, expressed in dB/decade?

Solution. This is still a scaling of ω^{-1}. A decade is a factor of 10, which means a factor of 10 reduction in amplitude, or $20\log_{10}(1/10) = -20\,\text{dB}$, so $-20\,\text{dB/decade}$.

2.3.7 Example Problem: High-Pass Filter

Consider the differentiator from Section 2.2.2.



In doing this problem you should see why this is also called a **high-pass filter**.

(a) Compute $\tilde{T}(\omega)$.
(b) Compute $T(\omega)$.
(c) Work out the low- and high-frequency asymptotics of $\tilde{T}(\omega)$.
(d) Find $f_{3\,\text{dB}}$.

Solution.
(a) Using the voltage-divider formula again,

$$\tilde{T}(\omega) = \frac{\tilde{V}_{\text{out}}}{\tilde{V}_{\text{in}}} = \frac{R}{R + X_C} = \frac{R}{R + i/\omega C} = \frac{\omega RC}{\omega RC + i}. \tag{2.48}$$

(b) Taking the modulus,

$$T(\omega) = \frac{\omega RC}{\sqrt{1 + (\omega RC)^2}}. \tag{2.49}$$

(c) Small ω:

$$\tilde{T}(\omega) \sim -i\omega RC. \tag{2.50}$$

Since $-i = e^{-i\pi/2}$, this *advances* the phase, and is a 90° **phase lead**, like the derivative $d/dt = -i\omega$ (hence, differentiator).

For large ω:

$$\tilde{T}(\omega) \approx 1. \tag{2.51}$$

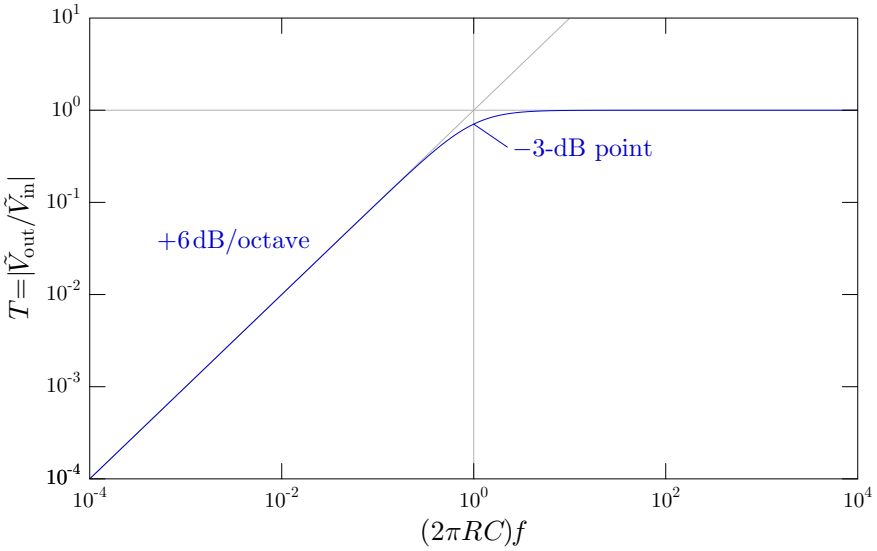
Hence, the high-pass filter.

(d) The 3-dB point occurs where $T(\omega) = 1/\sqrt{2}$, which has the same solution

$$f_{3\,\mathrm{dB}} = \frac{1}{2\pi RC} \tag{2.52}$$

as the low-pass filter.

Overall, the plot for the high-pass filter is basically a mirror image of the plot for the low-pass filter.



2.4 Phase

One thing that we haven't paid much attention to yet is the **phase shift** due to a linear circuit. (We did this just a bit, in looking at the low- and high-pass filters, in the asymptotic limits of low and high frequency, where the phase shift ended up being nothing or ±90°.) In general, the complex transfer function [from Eq. (2.43)],

$$\tilde{T}(\omega) := \frac{\tilde{V}_{\mathrm{out}}}{\tilde{V}_{\mathrm{in}}}, \tag{2.53}$$

gives information about both the amplitude [in $T(\omega)$] and phase (via the complex phase). In other words, we can always write

$$\tilde{T}(\omega) := T(\omega)\, e^{-i\phi(\omega)}, \tag{2.54}$$

where $\phi(\omega)$ is the frequency-dependent phase shift (remember the minus sign here is because of the phase convention in $e^{-i\omega t}$. If $\phi > 0$ at some frequency, we call this a **phase lead**, whereas $\phi < 0$ is a **phase lag**.

Now remember that for an arbitrary complex number z, we can write it in polar and cartesian forms as

$$z = r\, e^{-i\phi} = r\cos\phi - ir\sin\phi =: x + iy, \tag{2.55}$$

where r, ϕ, x, and y are all real. Equating real and imaginary parts, we get $x = r\cos\phi$ and $y = -r\sin\phi$, and dividing these equations gives

$$\frac{y}{x} = -\frac{\sin\phi}{\cos\phi} = -\tan\phi. \tag{2.56}$$

Solving for ϕ gives

$$\phi = -\tan^{-1}\frac{y}{x}, \tag{2.57}$$

remembering that $\tan x$ is an odd function. Now x and y are respectively the real and imaginary parts of $\tilde{T}(\omega)$, so

$$\phi(\omega) = -\tan^{-1}\left(\frac{\mathrm{Im}[\tilde{T}(\omega)]}{\mathrm{Re}[\tilde{T}(\omega)]}\right). \tag{2.58}$$
(phase shift of linear circuit)

Thus, we have the phase shift at any frequency in terms of the complex transfer function.

2.4.1 Example: Low-Pass Filter

In the low-pass filter, we had from Eq. (2.44)

$$\tilde{T}(\omega) = \frac{1}{1 - i\omega RC}. \tag{2.59}$$

Multiplying upstairs and downstairs by $1 + i\omega RC$, we can rewrite this as

$$\tilde{T}(\omega) = \frac{1 + i\omega RC}{1 + (\omega RC)^2}. \tag{2.60}$$

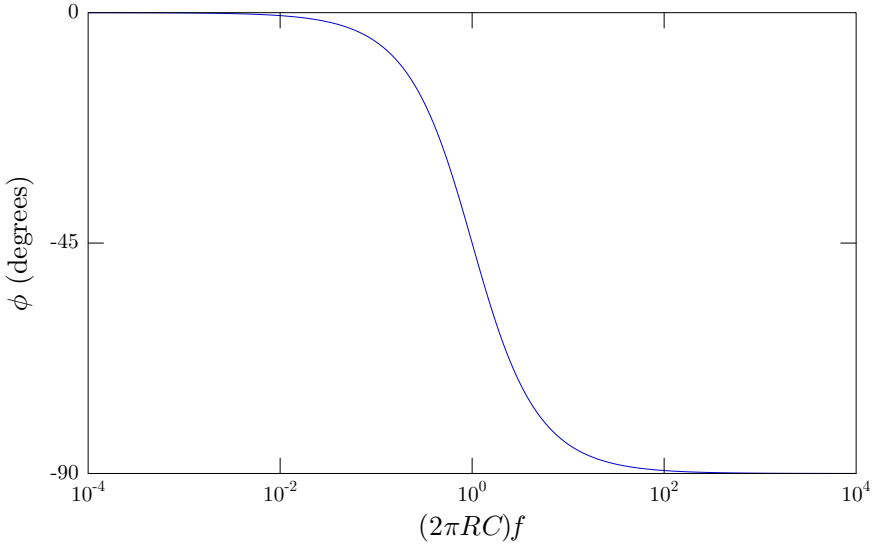
Now the real and imaginary parts are more obvious, and if we put these into Eq. (2.58), we get

$$\phi(\omega) = -\tan^{-1}(\omega RC). \tag{2.61}$$
(phase shift, low-pass filter)

How does this behave? In the extreme limits:

- For small frequencies [$\omega \ll (RC)^{-1}$], $\phi \approx -\omega RC$, which is a small (close to 0°) **phase lag**.
- For large frequencies [$\omega \gg (RC)^{-1}$], $\phi \approx -\pi/2$, which is a 90° **phase lag**.

In between, the phase lag moves smoothly between 0 and 90°, as shown in the plot below.



The high-pass filter is very similar, but the phase is 90° for small frequencies (in the "stop band"), and changes to 0° for large frequencies (in the "pass band").

2.5 Power

Now we arrive at the real meaning of why capacitors and inductors seem a lot like resistors, but with "imaginary resistance." Remember that the power dissipated in a circuit is $P = IV$. This is still true in an ac circuit, but

$$P(t) = I(t)\,V(t) \tag{2.62}$$

is the *instantaneous* power. This can either be positive or negative; positive means dissipating energy or perhaps storing energy in a capacitor or inductor, while negative means we are getting some stored energy back.

What we're interested in here is the *time-averaged* behavior, so we know the *net* effect of everything that happens over a cycle. So let's assume a monochromatic voltage and a current, with a possible phase shift ϕ in the current. In real notation, we have

$$\begin{aligned} V(t) &= V_0(\cos\omega t) \\ I(t) &= I_0\cos(\omega t + \phi). \end{aligned} \tag{2.63}$$

Then the power, time-averaged over one period $T = 2\pi/\omega$ is

$$\langle P\rangle = \frac{1}{T}\int_0^T V(t)\,I(t)\,dt. \tag{2.64}$$

Expanding the cosine in the current using the sum-angle formula,

$$I(t) = I_0[\cos(\omega t)\cos(\phi) - \sin(\omega t)\sin(\phi)] \tag{2.65}$$

and using this with the above expression for $V(t)$, we get

$$\langle P\rangle = \frac{I_0V_0}{T}\int_0^T [\cos^2(\omega t)\cos(\phi) - \cos(\omega t)\sin(\omega t)\sin(\phi)]\,dt. \tag{2.66}$$

In the second term the $\cos(\omega t)\sin(\omega t) = (1/2)\sin(2\omega t)$ averages to zero. In the first term $\cos^2(\omega t) = (1/2) + (1/2)\cos(2\omega t)$, which averages to just $1/2$. Thus,

$$\langle P\rangle = \frac{1}{2}I_0V_0\cos(\phi). \tag{2.67}$$

Now to simplify this a bit more, we want to compare this to time-averaged values of $V(t)$ and $I(t)$ separately. It doesn't make sense to time average them directly, because they average to zero. But we can compute the **rms** or **root-mean-square** values. This just means: square it, time-average it, take the square root, done. For the voltage, if we square it,

$$V^2(t) = {V_0}^2\cos^2\omega t, \tag{2.68}$$

then average it,

$$\frac{1}{T}\int_0^T V^2(t)\,dt = \frac{{V_0}^2}{2}, \tag{2.69}$$

then take the square root, we get the rms voltage:

$$V_{\text{rms}} = \frac{V_0}{\sqrt{2}}. \tag{2.70}$$
(rms voltage)

Similarly, for current,

$$I_{\text{rms}} = \frac{I_0}{\sqrt{2}}. \tag{2.71}$$
(rms current)

Note that these two rms expressions are valid *only* for sine waves. Then we can rewrite Eq. (2.67) as

$$\langle P \rangle = I_{\text{rms}} V_{\text{rms}} \cos(\phi). \tag{2.72}$$
(rms current)

It is common to define the **power factor** as

$$\text{power factor} := \frac{\langle P \rangle}{I_{\text{rms}} V_{\text{rms}}} = \cos(\phi). \tag{2.73}$$
(power factor)

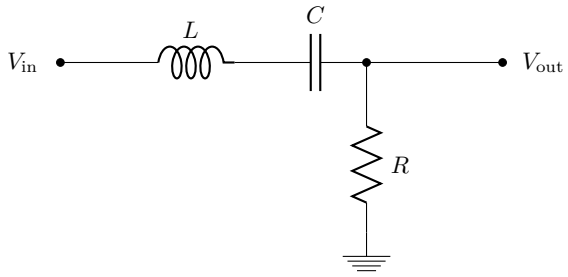
What does this mean? Breaking this down into cases:

- $\cos\phi = 1$ ($\phi = 0$): this occurs for a purely resistive load; the maximum power is dissipated here.
- $\cos\phi = 0$ ($|\phi| = \pi/2$): this is a purely reactive (capacitive or inductive) load, *no power is dissipated* (it is only stored and retrieved over each cycle).
- $0 < \cos\phi < 1$ ($0 < |\phi| < \pi/2$): there is some reactive component to the load impedance, so some power is dissipated, but not as much as for an equivalently large but real impedance.
- $\cos\phi < 0$ ($|\phi| > \pi/2$): it's also possible to have a *negative* power factor, which means the "load" is in fact a generator or EMF source.

Light bulbs and toasters are good examples of resistive loads. An example of a capacitive load is a piezo speaker or buzzer. Examples of inductive (reactive) loads are electric motors or lighting transformers or ballasts for fluorescent lights (the "magnetic" kind, not the "electronic" kind). Inductive loads are important in high-power applications, and the problem is that the power factor can be very small, so that large voltages and currents are needed to drive a motor. This isn't efficient, because the large voltages and currents cause wasted power to be dissipated *elsewhere*. A trick to help here is to "correct" the phase of the load impedance by connecting a parallel capacitor, which increases the power factor. It is common in air-conditioning compressor motors to have *two* capacitors, a "start" capacitor and a "run" capacitor. The start capacitor increases the parallel capacitance and hence the power factor when the motor first powers on, to give the motor extra startup torque.

2.6 Resonant Circuits

As another example of mixed impedances, we can consider a resonant filter with a resistor, inductor, and capacitor, as shown below.



Resonant LC circuits like this are also often called **tank circuits**. To analyze this, let's lump the series inductor and capacitor together into a single element, of impedance

$$\begin{aligned} Z_{LC} &= X_L + X_C = -i\omega L + \frac{i}{\omega C} \\ &= -\frac{iL}{\omega}\left(\omega^2 + \frac{1}{LC}\right) \\ &= -\frac{iL}{\omega}\left(\omega^2 - \omega_0^2\right), \end{aligned} \tag{2.74}$$

where we have defined the **resonant (LC) frequency**

$$\omega_0 := \frac{1}{\sqrt{LC}}. \tag{2.75}$$
(LC frequency)

Then we can treat what is left as a voltage divider. The transfer function is

$$\begin{aligned}\tilde{T}(\omega) &= \frac{R}{R + Z_{LC}} \\ &= \frac{R}{R - (iL/\omega)(\omega^2 - \omega_0^2)} \\ &= \frac{i\omega R/L}{(\omega^2 - \omega_0^2) + i\omega R/L}.\end{aligned} \tag{2.76}$$

Defining the damping constant

$$\gamma := \frac{R}{L}, \tag{2.77}$$

the transfer function becomes

$$\tilde{T}(\omega) = \frac{i\omega\gamma}{(\omega^2 - \omega_0^2) + i\omega\gamma}. \tag{2.78}$$
(transfer function, RLC circuit)

This is the same as the response function for a damped harmonic oscillator, with resonant frequency ω_0 and damping rate γ.

$$\tilde{T}(\omega) = \frac{\omega\gamma}{\sqrt{(\omega^2 - \omega_0^2)^2 + \omega^2\gamma^2}}.$$
(amplitude transfer function, RLC circuit) (2.79)

Note that on resonance ($\omega = \omega_0$), $T(\omega = \omega_0) = 1$ [and in fact $\tilde{T}(\omega = \omega_0) = 1$, so the signal is transmitted without amplitude reduction or any phase shift. Away from resonance, $T(\omega) < 1$, leading to a transmission "peak" around ω_0.

2.6.1 *Q* Factor

A common way to quantify the width of the resonance peak is the ***Q* factor**. The idea is to find the $-3\,\mathrm{dB}$ points of the resonance peaks (compared to the peak) as a measure of the width. Thus, setting $T(\omega) = 1/\sqrt{2}$,

$$\frac{1}{\sqrt{2}} = \frac{\omega\gamma}{\sqrt{(\omega^2 - \omega_0^2)^2 + \omega^2\gamma^2}}, \tag{2.80}$$

we can square this and rearrange to find

$$(\omega^2 - \omega_0^2)^2 = \omega^2\gamma^2. \tag{2.81}$$

This is the square of a quadratic equations, with four solutions

$$\omega = \pm\sqrt{(\gamma/2)^2 + \omega_0^2} \pm \frac{\gamma}{2}. \tag{2.82}$$

We only want the positive solutions, because $T(\omega) \geq 0$, and $T(\omega)$ has the same sign as ω. Thus,

$$\omega_{3\mathrm{dB}} = \sqrt{(\gamma/2)^2 + \omega_0^2} \pm \frac{\gamma}{2}. \tag{2.83}$$

If we define the width of the peak to be the difference between the -3-dB points,

$$\delta\omega_{3\mathrm{dB}} = \gamma. \tag{2.84}$$
(full width at half maximum)

This is also called the **full width at half maximum (FWHM)** ("half" here refers to the *power* transmission $|T(\omega)|^2$).

$$\omega^2 - \omega_0^2 = \omega\gamma. \tag{2.85}$$
(amplitude transfer function, RLC circuit)

Then we can write down the Q factor as the ratio of the resonance frequency to the FWHM:

$$Q = \frac{\omega_0}{\delta\omega_{3\text{dB}}} = \frac{\omega_0}{\gamma}. \tag{2.86}$$
(Q factor, RLC circuit)

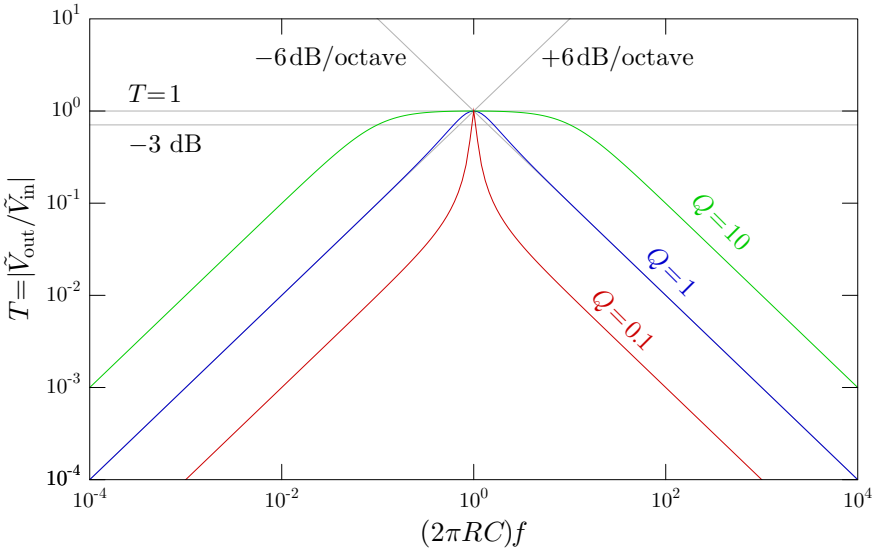
The "Q" here indicates the "quality" of the resonator: a large Q means a small γ compared to ω_0, and thus a narrow resonance. Actually, this is *not* the definition of the Q factor, but it's the one that physicists use when dealing with resonances—we'll come back to the formal definition below.

Also, it is useful to write the Q factor in terms of the electronic parameters. Using Eqs. (2.75) and (2.77),

$$Q = \frac{\sqrt{L}}{R\sqrt{C}}. \tag{2.87}$$
(Q factor, RLC circuit)

Thus, for example, an increased capacitance not only lowers the resonance frequency, but also the Q factor (because the Q factor depends on the resonance frequency).

The transfer function $T(\omega)$ is plotted below for three different values of Q.



Note that asymptotically, the filter rises/falls at ±6 dB/octave, like the low- and high-pass filters.

2.6.1.1 Fundamental Definition

Now to justify the simple formula (2.86) for the Q factor that we used. The more fundamental definition of the Q factor is defined in terms of stored energy in the oscillating circuit and the energy dissipated as follows:

$$Q := 2\pi \cdot \frac{\text{(maximum) stored energy}}{\text{energy loss per oscillation cycle}}. \tag{2.88}$$
(definition of Q factor)

The capacitor stored energy is, on average,

$$\langle E_C \rangle = \left\langle \frac{1}{2}CV^2 \right\rangle = \frac{1}{2}CV_{\text{rms}}^2, \tag{2.89}$$

while the inductor's stored energy is

$$\langle E_L \rangle = \left\langle \frac{1}{2} L I^2 \right\rangle = \frac{1}{2} L I_{\rm rms}^2. \tag{2.90}$$

These expressions are actually the same; using $I = C(dV/dt)$ and time-averaging gives $I_{\rm rms}^2 = \omega^2 C^2 V_{\rm rms}^2$, and so, for example,

$$\langle E_C \rangle = \frac{1}{2} C V_{\rm rms}^2 = \frac{1}{2\omega_0^2 C} I_{\rm rms}^2 = \frac{1}{2} L I_{\rm rms}^2, \tag{2.91}$$

where we used $\omega^2 = \omega_0^2 = 1/LC$. The power dissipated is

$$\langle P \rangle = I_{\rm rms} V_{\rm rms}, \tag{2.92}$$

and so the energy lost per cycle is

$$T\langle P \rangle = T I_{\rm rms} V_{\rm rms}, \tag{2.93}$$

where T is the period. Putting this together,

$$Q = 2\pi \left(\frac{L I_{\rm rms}^2}{T I_{\rm rms} V_{\rm rms}} \right) = \omega_0 \left(\frac{L I_{\rm rms}}{V_{\rm rms}} \right) = \omega_0 \frac{L}{R} = \frac{\omega_0}{\gamma}, \tag{2.94}$$

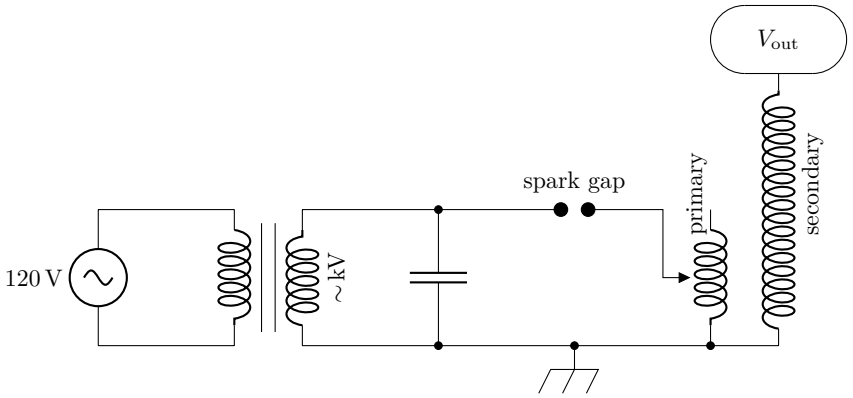
in agreement with Eq. (2.86).

2.7 Circuit Practice

2.7.1 Tesla Coil

A nice example of a resonant circuit is the **Tesla coil**. There are many variations, but the simplest circuit is below. The idea is to put in a fairly "low" voltage (line voltage), and generate an output voltage as large as possible, preferably at least hundreds of kV or even MV.

Here we will go through a basic tesla coil circuit to give you some experience in "reading" a more complex circuit diagram.



Typical designs use a neon-sign transformer in the first stage to boost the voltage to ∼6 kV or more. This drives an RLC circuit (the "R" is the wire resistance), which oscillates at a tunable resonant frequency (by tuning the variable inductor). A spark gap also interrupts the RLC circuit; the nonlinear sparking increases dI/dt to help get a larger output voltage in the secondary transformer. The secondary coil is just a long air-core coil with many turns, with a large electrode on top. The capacitance of the secondary circuit is the capacitance due to the windings of the coil, as well as the output terminal (the other electrode being ground). So the secondary circuit is also an RLC oscillator. The primary is tuned to resonate with the secondary. On resonance, the voltage multiplier is *not* the ratio of turns as in a normal transformer, but rather it turns out to be the ratio of the Q factors of the two RLC oscillators. The oscillators are generally tuned to radio frequencies (hundreds of kHz), which makes the output (relatively) safe for humans.

2.8 Exercises

Problem 2.1

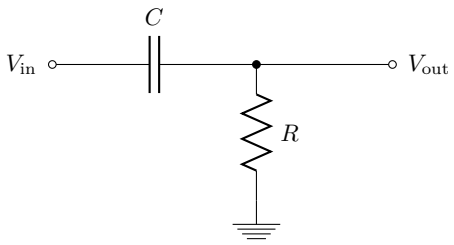
Use the capacitor law $Q = CV$ to show that the effective capacitance of two *parallel* capacitors is $C_{\text{eff}} = C_1 + C_2$, where C_1 and C_2 are the capacitances of the individual capacitors.

Problem 2.2

Show that the effective capacitance C_{eff} of two capacitors C_1 and C_2 in *series* is given by $C_{\text{eff}}^{-1} = C_1^{-1} + C_2^{-1}$.

Problem 2.3

Consider the differentiator circuit shown below, where $V_{\text{in}}(t)$ is an arbitrary input voltage.



(a) Show that the differential equation for this circuit is given by

$$\frac{dV_{\text{out}}}{dt} = -\frac{V_{\text{out}}}{RC} + \frac{dV_{\text{in}}}{dt}. \tag{2.95}$$

(We did this in class.)

(b) Use the integrating-factor trick that we used for the integrator circuit (i.e., define $\tilde{V} := V_{\text{out}} e^{t/RC}$, simplify the equation, and integrate from 0 to t) to find the general solution

$$V_{\text{out}}(t) = V_{\text{in}}(t) + \big[V_{\text{out}}(0) - V_{\text{in}}(0)\big] e^{-t/RC} - \frac{e^{-t/RC}}{RC} \int_0^t V_{\text{in}}(t')\, e^{t'/RC}\, dt'. \tag{2.96}$$

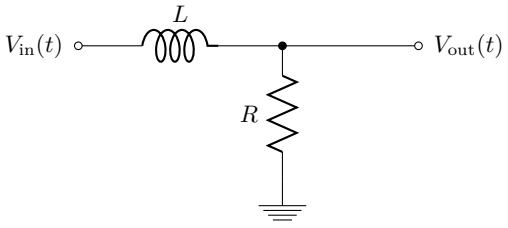
(c) Write down the solution in the case where V_{in} is turned on suddenly to a constant value from 0, *just after* $t = 0$. Give a brief, qualitative description of the solution (use a sketch if you need to).

Note: the reason for assuming the turn-on comes just after $t = 0$ is a bit technical, but it's necessary to get the boundary terms in the general solution to come out right. Basically, the "leading edge" of the input signal is an "interesting" part of the signal (the most interesting, actually), and so we must make sure to capture it in our interval of integration. Equivalently, we could have taken out integration range from $-\infty$ to t, taking our boundary condition to be $V_{\text{in}}(-\infty) = 0$.

Another note: From your solution here, you should see the reason why this circuit is called a "dc block."

Problem 2.4

Consider the LR circuit below.



(a) Is this a high-pass or a low-pass filter? Give a *qualitative* argument for your answer.

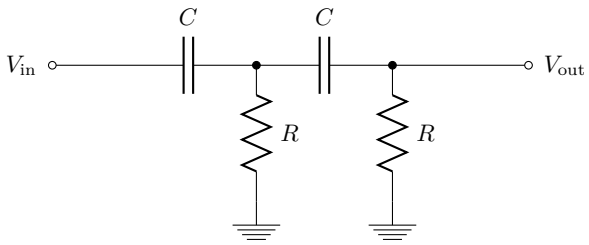
(b) Write down a differential equation relating $V_{\text{in}}(t)$ and $V_{\text{out}}(t)$.

(c) Solve the equation for an arbitrary input $V_{\text{in}}(t)$ (*not* necessarily a single frequency!), by using the integrating-factor method. You should end up with a solution in the form of an integral over $V_{\text{in}}(t)$. (*Hint*: it may help to define $\tilde{V} := Ve^{(R/L)t}$ for one of the voltages.

(d) Write down the solution for the case where the input voltage is turned on suddenly from 0 to V_0 *just after* $t = 0$, and then held at V_0 for all $t > 0$. You can assume the input was zero for all times in the past. What is the time constant of your solution?

Problem 2.5

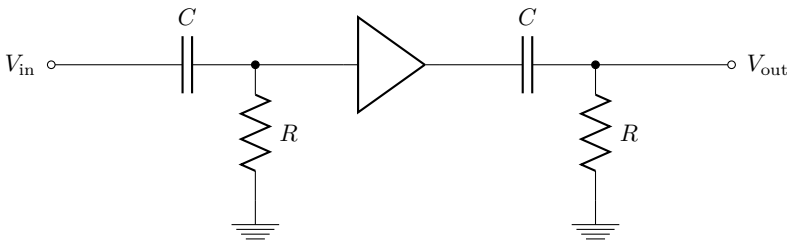
Consider the circuit shown below of two cascaded high-pass filters.



Find the (complex) transfer function $\tilde{V}_{\text{out}}/\tilde{V}_{\text{in}}$, assuming an input frequency of ω. How does this scale for large and small frequency? (Also give the scaling at small frequencies in dB/octave.)

Problem 2.6

Consider the circuit shown below of two cascaded high-pass filters, separated by a **buffer amplifier**.



The buffer amplifier has two important properties: the input (left-hand-side connection) draws no current (and thus produces no load on the first RC filter), and the output voltage is equal to the voltage at the input.

(a) Find the (complex) transfer function $\tilde{V}_{\text{out}}/\tilde{V}_{\text{in}}$, assuming an input frequency of ω. Examine the scaling behavior for large and small frequencies, and give the scaling at small frequencies in dB/octave.

(b) Make a (log-log) plot of the amplitude transfer function $T(\omega) = |\tilde{V}_{\text{out}}(\omega)/\tilde{V}_{\text{in}}(\omega)|$. Also include on the same plot the corresponding transfer functions from the Problem 5, and the transfer function for a simple high-pass filter.

(c) Make another plot of the phase of the output compared to the input for the same three circuits as in (b). That is, if we write out the amplitude and phase of the transfer function as

$$\tilde{T}(\omega) = T(\omega)\, e^{-i\phi}, \tag{2.97}$$

then make a plot of ϕ vs. ω. Here, use a logarithmic frequency axis, and a linear phase axis. Be clear about the nature of the phase shift (lead vs. lag).

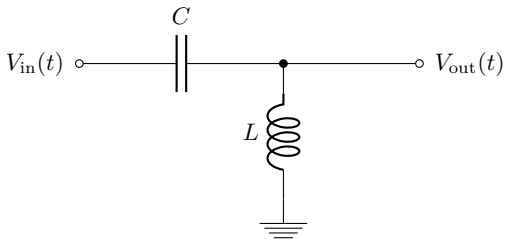
Note: use any program you like for plotting. For a plotting tutorial in Mathematica, see

`http://atomoptics.uoregon.edu/~dsteck/teaching/mathematica/plot-tutorial.nb`
`http://atomoptics.uoregon.edu/~dsteck/teaching/mathematica/plot-tutorial.pdf`

In Mathematica, use `LogLogPlot` for a log-log plot and `LogLinearPlot` for a log plot in the x-direction instead of `Plot`.

Problem 2.7

Consider the circuit below, consisting of a capacitor of capacitance C, and an inductor of inductance L.



(a) Is this a high-pass or a low-pass filter? Give a *qualitative* argument for your answer.

(b) Derive the transfer function $[\tilde{T}(\omega)]$ for the circuit.

(c) Derive the amplitude transfer function $[T(\omega)]$ for the circuit.

(d) Give the scaling of the transfer function in the "stop band" of the circuit [i.e., the asymptotic region where $T(\omega)$ is small]. Express your answer in dB/octave.

Problem 2.8

A "schmapacitor" of "schmapacitance" S is a bipolar component whose operation is defined by the relation

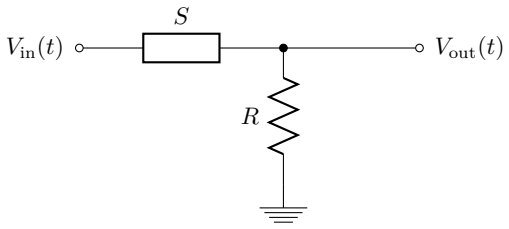
$$I = S\frac{d^3V}{dt^3}, \tag{2.98}$$

where I is the schmapacitor current and V is the voltage drop across the schmapacitor.

Derive expressions for the effective schmapacitance of two schmapacitors in series and for two schmapacitors in parallel.

Problem 2.9

Consider the circuit below, consisting of a resistor of resistance R, and a "schmapacitor" of "schmapacitance" S.



The schmapacitor is defined as in Problem 8 by the relation

$$I = S\frac{d^3V}{dt^3}, \tag{2.99}$$

where I is the schmapacitor current and V is the voltage drop across the schmapacitor.

(a) Is this a high-pass or a low-pass filter? Give a *qualitative* argument for your answer.

(b) Derive the transfer function $[\tilde{T}(\omega)]$ for the circuit.

(c) Derive the amplitude transfer function $[T(\omega)]$ for the circuit.

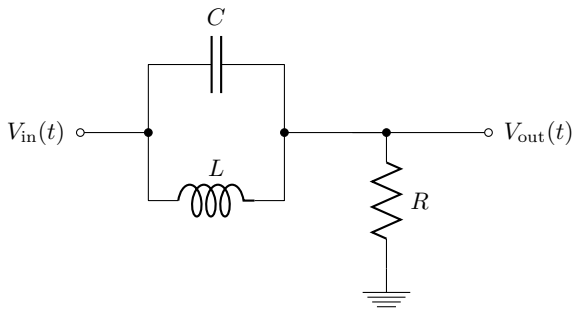
(d) Give the scaling of the transfer function in the "stop band" of the circuit [i.e., the asymptotic region where $T(\omega)$ is small]. Express your answer in dB/octave.

(e) What is the asymptotic phase in the stop band? Is it a phase lead or a phase lag?

(f) Derive an expression for the -3-dB frequency $f_{3\text{dB}}$ of the circuit.

Problem 2.10

Consider the RLC circuit below.



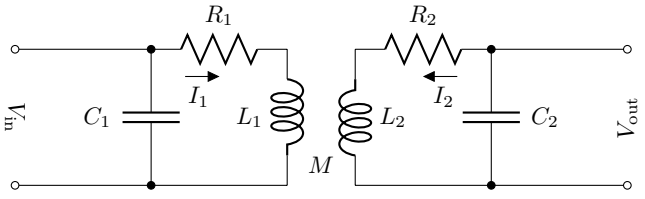
(a) Derive the transfer function $\tilde{T}(\omega)$ for the circuit.

(b) Derive the amplitude transfer function $[T(\omega)]$ for the circuit.

(c) What is the behavior of $\tilde{T}(\omega)$ for small and large frequencies? What happens at $\omega = 1/\sqrt{LC}$? Based on your answers, give a *qualitative* description of how this circuit functions as a frequency-dependent filter.

Problem 2.11

Consider the circuit below, consisting of two coupled, resonant circuits. In the resonant case that we will analyze, this circuit is useful for boosting the amplitude of ac signals, similar to ordinary transformer circuits. This is also the key to how the Tesla coil works.



The two inductors are coupled by a **mutual inductance** M, such that, for example, the voltage drop across the inductor is

$$V_{L_1} = L_1 \frac{dI_1}{dt} + M \frac{dI_2}{dt}. \tag{2.100}$$

Let q_1 and q_2 be the charges on capacitors C_1 and C_2, respectively (we will reserve Q_1 and Q_2 for the Q factors of each circuit later).

(a) Derive a coupled pair of differential equations for q_1 and q_2. One of the equations should look like

$$\omega_1^2 q_1 + \gamma_1 \dot{q}_1 + \ddot{q}_1 + \frac{M}{L_1} \ddot{q}_2 = 0, \tag{2.101}$$

where the resonant frequencies of the circuits are

$$\omega_{1,2} := \frac{1}{\sqrt{L_{1,2}C_{1,2}}}, \tag{2.102}$$

and the damping rates are

$$\gamma_{1,2} := \frac{R_{1,2}}{L_{1,2}}. \tag{2.103}$$

(b) Now we will consider a sinusoidal input signal at frequency ω. The differential equations from (a) also apply to complex charges $\tilde{q}_{1,2}(t)$, with time dependence of the form

$$\tilde{q}_{1,2}(t) = q_{1,2}\, e^{-i\omega t}. \tag{2.104}$$

Use this to eliminate the derivatives in the coupled ODE's from (a), to write a pair of algebraic equations for the complex *amplitudes* $q_{1,2}$.

(c) Write your equations from (b) as a matrix equation, in the form

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = 0. \tag{2.105}$$

Note that for this to hold, the determinant of the $2x2$ matrix must vanish; i.e., $AD - BC = 0$. Use this to derive the condition

$$(\omega_1^2 - \omega^2 - i\omega\gamma_1)(\omega_2^2 - \omega^2 - i\omega\gamma_2) = \omega^4 k^2, \tag{2.106}$$

where

$$k := \frac{M}{\sqrt{L_1 L_2}} \tag{2.107}$$

is the **coupling coefficient** for the inductor pair (satisfying $0 \le k \le 1$). This is a quartic equation in ω; the physical solutions determine the resonant frequencies of the coupled-oscillator system. For now, note that in the limit of small coupling k and small damping $\gamma_{1,2}$ (which we will assume from now on), the solutions to this equation are $\omega = \pm\omega_0$ (note that the frequencies will appear as positive/negative pairs; the simplest interpretation is that the positive solution is physical).

(d) The first regime we will analyze is the ideal-resonator limit, where $R_{1,2} = 0$ (and thus $\gamma_{1,2} = 0$), for matched resonators ($\omega_1 = \omega_2 =: \omega_0$). Derive the resonant frequencies from Eq. (2.106). Then use your solution from part (b) to derive an expression for q_2/q_1, and finally derive an expression for the transfer function $\tilde{V}_{\rm out}/\tilde{V}_{\rm in}$, which you should find to be independent of frequency. You should find that the result involves the ratio of the number of windings on the two inductors, as expected for an ordinary transformer (you may use $L \propto N^2$, where N is the number of windings on the inductor).

(e) Finally, we will examine the case of matched resonators ($\omega_1 = \omega_2 =: \omega_0$), with small but finite damping ($\gamma_{1,2} \neq 0$, $\gamma_{1,2} \ll \omega_0$). Assume the input is driven on resonance ($\omega = \omega_0$), and derive an expression for the transfer function $\tilde{V}_{\rm out}/\tilde{V}_{\rm in}$. You should find that your answer involves the ratio of the Q factors

$$Q_{1,2} := \frac{\omega_{1,2}}{\gamma_{1,2}}, \tag{2.108}$$

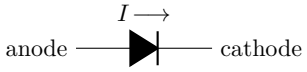
as well as the ratio of the number of turns.

Chapter 3

Diodes

3.1 Ideal Diode

Diodes are useful circuit elements: in the simplest sense, they act as one-way valves or "trapdoors" for electrons and thus for electrical current. Schematically, the following symbol represents them:



The two terminals are called the **anode** and **cathode**. The rules that govern the ideal diode are as follows:

1. If $V_{\text{anode}} > V_{\text{cathode}}$, then the diode acts like a **short circuit**: lots of current can flow. In this case, the diode is said to be **forward-biased**.

2. If $V_{\text{anode}} \leq V_{\text{cathode}}$, then the diode acts like an **open circuit**: no current flows at all. In this case, the diode is said to be **reverse-biased**.

Of course, the situation with real diodes is more complicated, and we'll get to that. For now, a good mnemonic to remember the direction of current flow is that the diode symbol makes an "arrow" in the direction of the current flow (towards the cathode). Also, in the lab, a typical "signal diode" (for mA-level currents) looks schematically like this:

There is usually a band (not always dark) that marks the cathode end; you can think of the band as being a "minus sign" that marks the cathode.

The diode is obviously a *nonlinear* device, since voltage is not simply proportional to current. Since Ohm's law $V = IR$ does not hold in a simple way with constant R, so a diode is an example of a **non-Ohmic** device.

3.2 Vacuum Diodes

The name of the diode comes from the original vacuum-tube realization of a diode, which has two electrodes (hence, the "di"). The cathode is a heated electrode that "boils off" electrons; whether the electrons make it to the anode—and hence whether current flows—is determined by the relative voltage on the anode, because the anode with either repel or attract the electrons from the cathode. (Think about it to see that the current flows only with the correct voltage polarity.)

3.3 Semiconductor Diodes

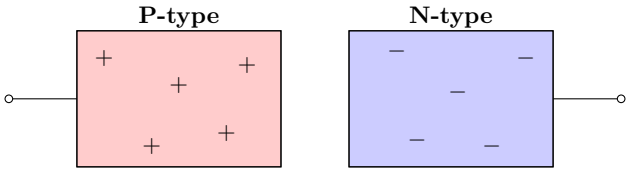
By far the most important realization of a diode is using semiconductor materials, so that's what we'll concentrate on here. Roughly speaking, a *semiconductor* is a material somewhere in between a conductor and an insulator. In an insulator, electrons, are bound in place, so they can't move around to form a current. In a conductor, electrons move freely, and current flows (there is little resistance). In a semiconductor, the electrons are *mostly* bound, but a few are thermally activated into conducting states, so conduction can happen.

But more important are **doped** semiconductors, where impurities are added to enhance conduction. There are two types:

1. In **n-type semiconductors**, the semiconductor is doped with impurities that introduce *excess* electrons (called **n-type carriers**, "n" for for "negatively charged"). The n-type carriers do the conducting.

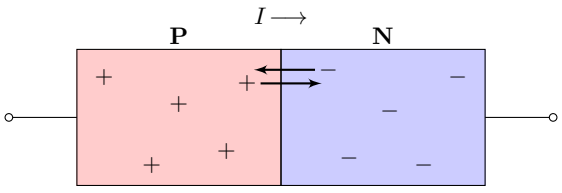
2. In **p-type semiconductors**, the semiconductor is doped with impurities that introduce a *deficit* of electrons. The *absence* of an electron is something like a positive charge, and is called an **electron hole**, or just **hole**. (These are then called **p-type carriers**, "p" for for "positively charged"). The p-type carriers (holes) do the conducting here.

A **semiconductor diode** is a **junction** between p-type and n-type semiconductors (called a **p-n junction**). If we first consider *separate* p- and n-type semiconductors, we get something like this:



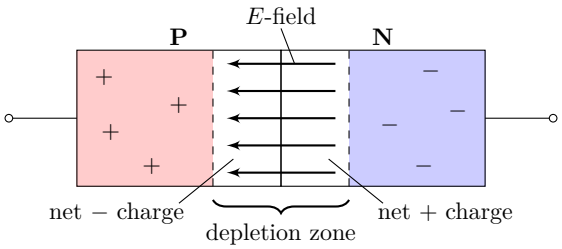
Note that despite the presence of the carriers in each semiconductor, the semiconductors are *electrically neutral*.

Now let's smoosh these together and see what happens. There is now a p-n junction, the charges are free to diffuse across due to random, thermal motion.



(no depletion zone)

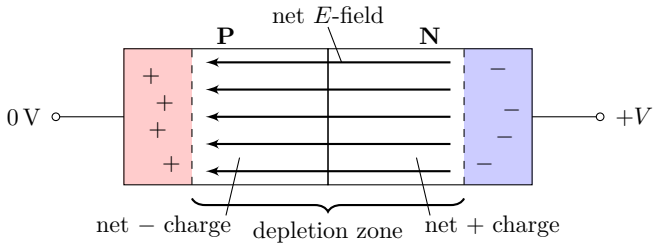
Remember that the semiconductors were neutral *before* we put them together, so now that carriers are moving across, this sets up net charges on either side of the junction, which creates an electric field. All this is summarized in the diagram below.



Note that for any carriers that diffused across the junction, the electric field tries to "restore" them back to their original home. So the diffusion continues until the electric field builds up to the point where just balances the tendency for more carriers to diffuse across the junction. Meanwhile, the n-type carriers that make it across to the p-type material will **annihilate** the p-type carriers (electron + hole = nothing), and the same thing happens for p-type carriers that make it across the junction to the n-type material. So there is a region around the junction called the **depletion zone**, where there are *no* carriers. Note that in this equilibrium state, no net current can flow, because there are no carriers to transport charge (current) across the depletion zone.

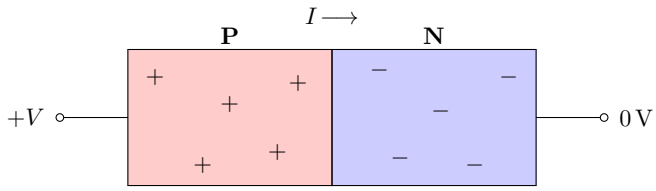
Now let's argue that this p-n junction realizes the diode as shown below.

First, let's do the **reverse-biased case**, where no current should flow. That is, suppose we set the anode to 0 V, and bring the cathode up by $+V$ relative to the anode. Then the situation is shown in the diagram below.



In this case, the electric field due to the applied voltage *adds* with the electric field from the carrier diffusion. This also pulls the p-type carriers towards the anode (the terminal on the p-type material), and n-type carriers towards the cathode. The net effect is just to *increase* the width of the depletion zone.

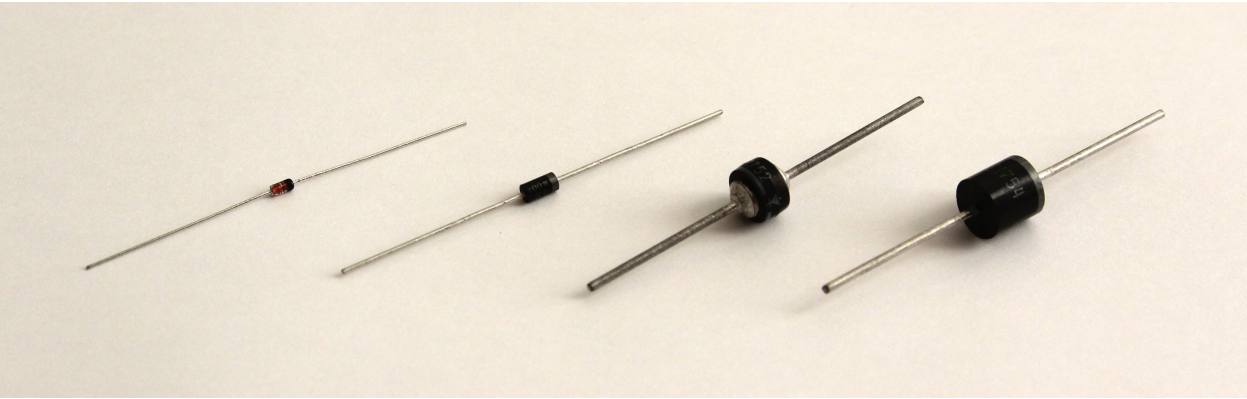
Now let's do the **forward-biased case**, where the diode should conduct. So we bring the anode to a voltage $+V$ relative to the cathode. Now the external field due to this potential difference points from left to right, and *cancels* the electric field due to carrier diffusion, and it tries to make the carriers in each material move towards the junction. The net effect is to *shrink* the depletion zone. Once a *sufficiently strong voltage* is in place, the depletion zone completely disappears, and a net current can flow, as shown below.



(no depletion zone)

Note that the current here involves each type of carrier moving towards the junction, where it annihilates one of the opposite type.

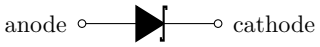
A few typical diodes are shown in the photo below.



From left to right, these are: 1N914B (signal diode), 1N4001 (1-A, general-purpose rectifier), MR752 (6-A, 200-V rectifier), GI754 (6-A, 400-V rectifier). All cathodes point to the upper-right-hand corner.

3.3.1 Schottky Diodes

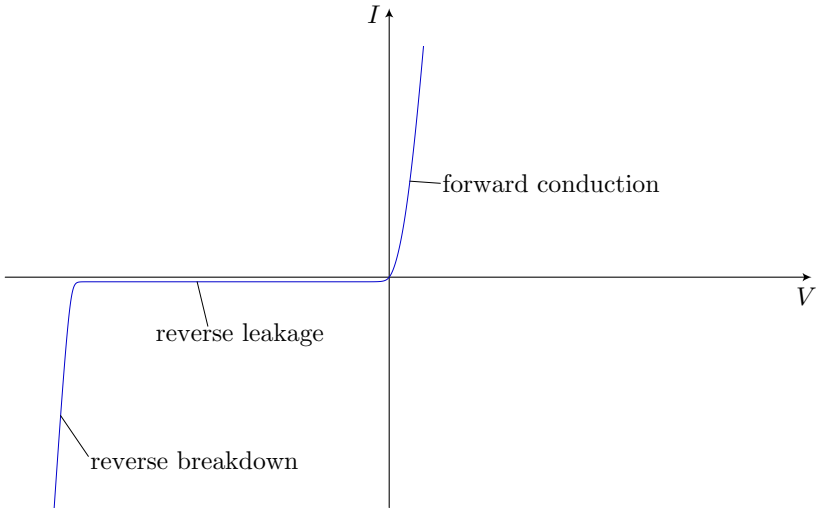
A variation on the semiconductor diode is the **Schottky diode**, which replaces the p-type material with a conductor. The metal-semiconductor junction is called a **Schottky barrier**. The operation is similar, the advantages being faster switching between forward- and reverse-biased modes, and lower forward voltage drop (as we will discuss shortly). The Schottky diode has a modified symbol to distinguish it from a "regular" diode, as shown below.



However, it is functionally the same as a semiconductor diode.

3.4 Current–Voltage Characteristics

All these effects lead to a more complicated relation between voltage and current. Below is a schematic plot of the V–I relation for a diode.



This is the *forward* current I plotted against the *forward* voltage V; *negative* values of course indicate that the voltage/current are going in the reverse-biased direction. There are a few features to notice here.

1. **Forward conduction.** As the diode is forward-biased, the forward current rapidly (in fact, exponentially) increases. But for any forward current, there is a **forward voltage drop**—remember this is

needed to squish the depletion region down to nothing so the diode can conduct. A handy number to remember for quick calculations is 0.7 V for the voltage drop, for forward-biased silicon diodes. The drop is somewhat more for high-power diodes, somewhat less for Schottky and germanium diodes. For example, the common 1N914 (silicon signal diode) has a forward voltage drop of 0.7 V at 10 mA, dropping to 0.6 V at 1 mA and going up to 0.9 V at 100 mA. The 1N5711 Schottky diode (small-signal diode, rated for 15-mA maximum current) has a similar drop of about 0.7 V at 10 mA, but a comparatively lower drop of 0.4 V at 1 mA. These numbers are all temperature-dependent, but the values here are at 25°C, and vary from device to device.

2. **Reverse leakage.** When the diode is reverse-biased, the current is not *completely* blocked, but some current flows. This is roughly constant over a wide range of reverse-bias voltages, and is usually labelled I_S, with the "S" for **saturation current**. This is also called, more obviously, the **reverse-leakage current**. This is typically ∼10 nA for silicon and Schottky diodes.

3. **Reverse breakdown.** If the reverse-bias voltage is sufficiently large, the insulating properties of the diode break down (like any insulator), and the diode conducts. The voltage at which the diode starts to conduct is the **reverse-breakdown voltage**, and is at least 100 V for the 1N914 and at least 70 V for the 1N5711.

3.4.1 Diode Law

Neglecting the reverse breakdown, the diode V–I relation is reasonably well-described by the **diode law**,

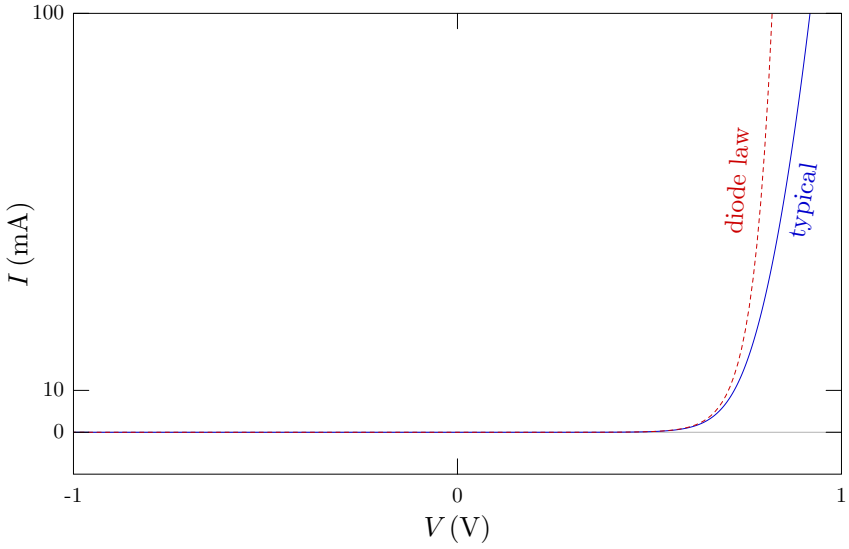
$$I = I_S \left(e^{V/nV_T} - 1 \right), \tag{3.1}$$
(diode law)

where I_S is the saturation current that we already discussed, V_T is the **thermal voltage**

$$V_T := \frac{k_B T}{e}, \tag{3.2}$$
(diode law)

$k_B = 1.381 \times 10^{-23}$ J/K is the **Boltzmann constant**, $e = 1.602 \times 10^{-19}$ C is the **fundamental charge** (magnitude of the electron charge), and T is the absolute temperature. The thermal voltage is 25.69 mV at 25°C. Also, n is the **ideality factor**, which typically falls in the range of 1 to 2, and is a sort of "fudge factor" for real junctions. Often this is just set to 1 for simplicity, in which case the diode law becomes the "ideal diode law."

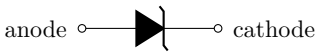
Note that the diode law is only really valid below the "knee" of the V–I curve, before the current starts to really take off. As an example, consider the plot below of two models for the 1N914 at 25°C.



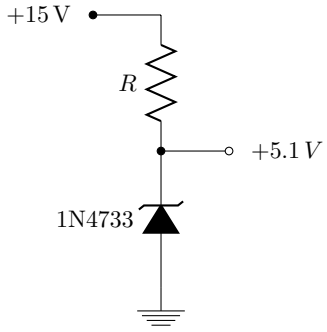
The "diode law" (dashed) curve shows the diode law with $I_S = 6.2229 \times 10^{-9}$ and $n = 1.9224$. The "typical" (solid) curve shows a more complete diode model.[1] The simple model overestimates the conducted current at larger forward voltages.

3.5 Zener Diodes

A **Zener diode** (pronounced ZEE-ner) is a regular diode with a carefully engineered reverse-breakdown voltage, typically in the range of 3–100 V. The Zener diode has a special symbol, as shown below.

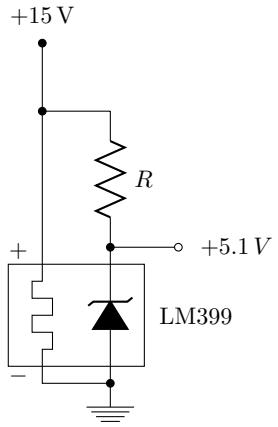


The Zener diode is mainly useful as a voltage regulator. A typical circuit is shown below.



The resistance R here depends on the intended load. The idea is that the Zener diode "wants" to operate at the reverse-breakdown voltage (5.1 V for the 1N4733), and it draws just the right amount of current to make the voltage drop across R equal to the difference between the supply (+15 V) and output (+5.1 V) voltages. How does the Zener diode "know" how much current to draw? You should try thinking this through: Think of the reverse-biased diode as a variable resistor, and think of the circuit as a voltage divider. The diode has large resistance for small currents, and small resistance for large currents. The only self-consistent point is for the diode to drop the breakdown voltage, which is the transition between these two regimes.

The main problem with this circuit is that, as we have seen, the properties of diodes depend on temperature. A good solution when you need a precise voltage reference is a **temperature-stabilized Zener diode**, like the LM399, which has a breakdown voltage of 6.95 V, and an integrated oven to keep the Zener's temperature constant.



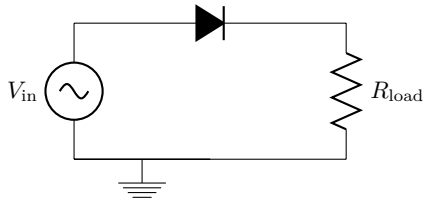
[1] This is the SPICE (circuit-simulator) model for the 1N914 from Central Semiconductor; the I_S and n values in the diode law are the same as the parameters in the SPICE model.

The extra two connections are the power-supply leads for the oven heater. Note that this circuit can draw large (up to 200 mA) for the first few seconds after the circuit is turned on, while the oven temperature stabilizes. You can then obtain other reference voltages using a voltage divider or an op-amp circuit to multiply the voltage by a known factor—something we will get to later.

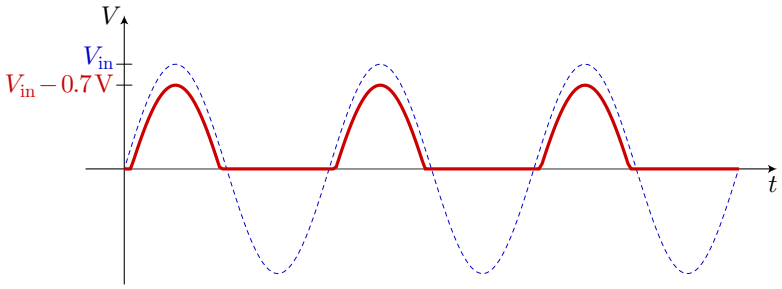
3.6 Rectifier Circuits

3.6.1 Half-Wave Rectifier

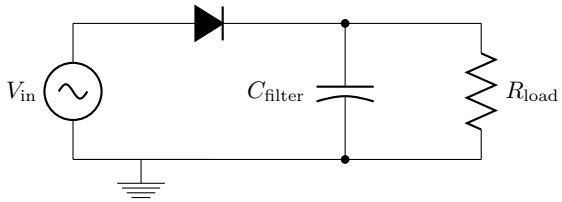
One of the main uses of a diode is as a **rectifier**, or something that changes an alternating-current (ac) signal into a direct-current (dc) signal. This is especially useful for changing the line voltage (120 V in the U.S.) into a dc voltage (e.g., for a power supply), after stepping the line voltage down (or up) by some factor using a transformer. The simplest example of a rectifier circuit is the **half-wave rectifier**, which uses only a single diode.



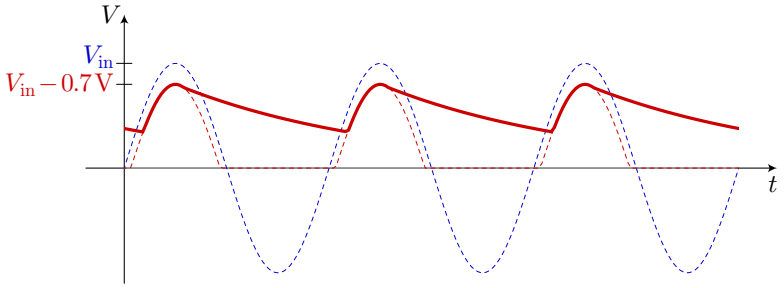
Here, the supply (with *amplitude* V_{in}) creates a voltage across the load (represented schematically here by a load resistor R_{load}); since current only flows in one direction, the supply can only impose a positive voltage across the resistor, and the voltage during forward conduction is always a diode drop *below* the voltage of the power supply.



This is not very much like a dc voltage, and not a very good dc power supply. The trick to getting a better dc signal is to add a smoothing capacitor across the load, as in the schematic below.



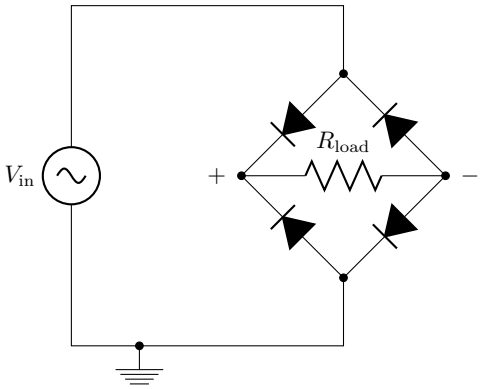
Then the circuit charges the capacitor through the diode to a maximum of V_{in}, less the diode drop. When the rectified voltage falls away, the capacitor "props up" the output voltage, which decays exponentially with a $1/e$ time of $R_{\text{load}}C_{\text{filter}}$. The net result is shown below.



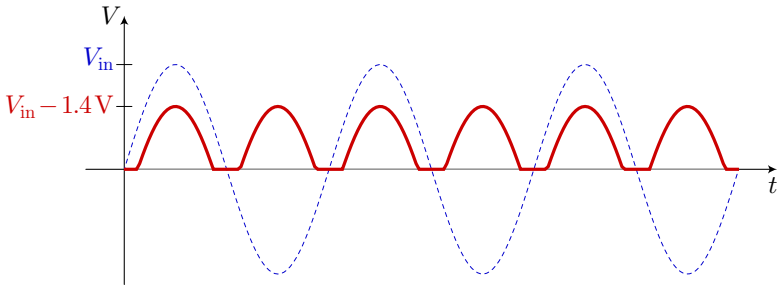
The resulting wave has **ripple**, and choosing the RC time to be as long as possible makes the ripple very small (for good power supplies, the ripple should be of order mV or tens of mV, depending on the current and application). Smaller load resistances (i.e., loads drawing more current) produce more ripple, while larger smoothing capacitors reduce the ripple.

3.6.2 Full-Wave Rectifier

A major problem with the half-wave rectifier is the long time between rectified peaks, making it difficult to get small ripple. A solution to this is the **full-wave rectifier**, which uses four diodes as shown below.

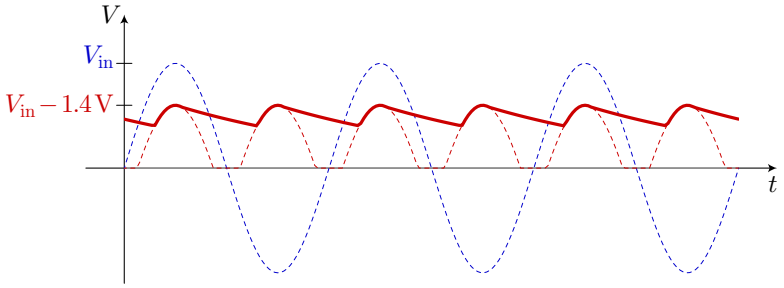


This passes both the positive and negative peaks, as shown below. Trace the current through the diodes on both phases of the input to convince yourself that this works out.



We don't "waste" the negative peaks here, but the price is that we lose *two* diode drops from the input voltage. Also, note that we are plotting the **voltage difference** across the load resistor; the absolute voltages on either end are more complicated, because the output voltage is referenced to the input voltage in a way that depends on which diodes are conducting at the moment. If the input voltage is **floating** (i.e., not referenced to ground, which is usually the case of a transformer output), we can instead ground the − side of the load, in which case the *absolute* output voltage goes from 0 to $V_{in} - 1.4\,V$.

Adding a capacitor to this setup, we get a decent, filtered power supply.

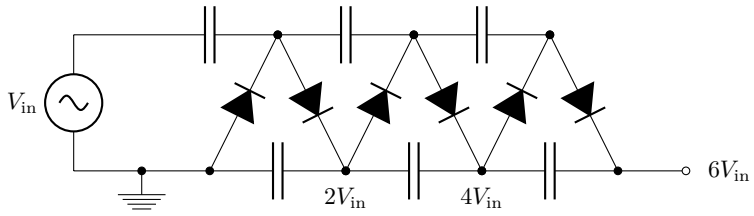


Again, we lose almost another volt compared to the half-wave rectifier, but for the same load and filter capacitor, we get much less ripple. Yet better power supplies (with smaller ripple) can be realized by adding a linear regulator IC.

3.7 Circuit Practice

3.7.1 Cockroft–Walton Multiplier

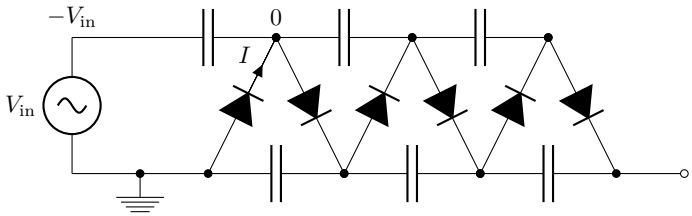
Explain how the circuit below, the **Cockcroft-Walton multiplier**, works. The multiplier is driven by an ac source of amplitude V_{in}, and each "stage" of the multiplier consists of two capacitors and two diodes. The output after N stages is (in steady state) a dc voltage of $2NV$. Three stages are shown in the example below.



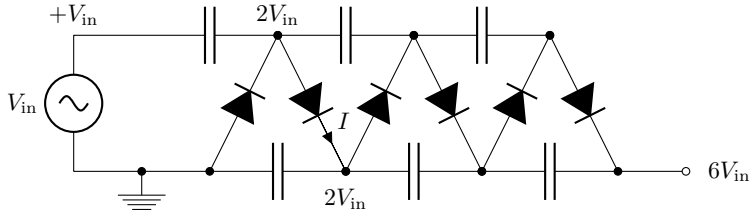
Assume ideal capacitors and diodes (and think about why it is reasonable to ignore the forward voltage drop of the diodes in this kind of circuit).

Solution. The idea in multiplying voltages is to get *high* voltages, so the input ac voltage would be of the order of 1 kV, in which case a diode drop of 0.7 V makes little difference.

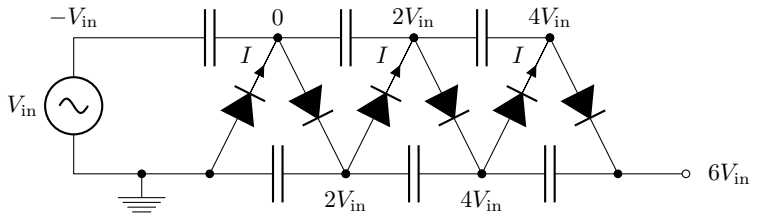
To analyze the circuit, we will first trace the voltages when the input voltage has swung low. Current flows across the first diode to charge the first capacitor.



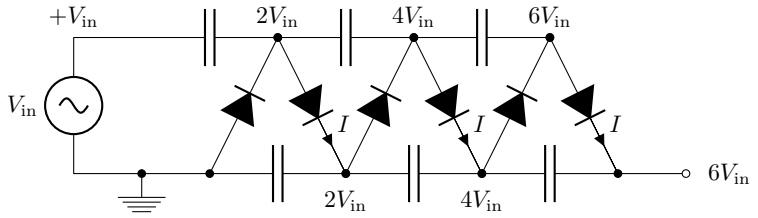
Then, on the positive-voltage phase, the absolute voltage of the first capacitor increases, but the voltage *across* it stays the same. Current flows to charge the second capacitor.



The process continues. In steady state, to balance any leakage of current from the output terminal, current would flow as follows in the negative-input phase,



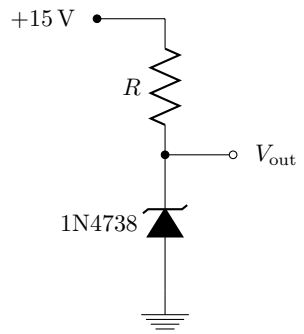
which switches to the following during the positive-input phase:



3.8 Exercises

Problem 3.1

Consider the zener-diode voltage regulator shown below.



Suppose you design this circuit to drive (supply) a load resistance R_L.

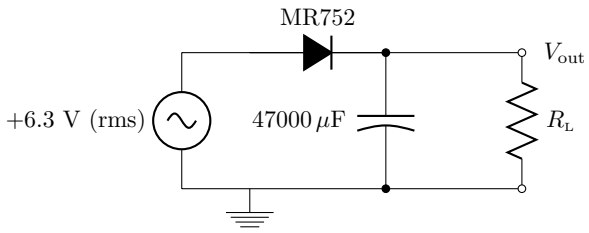
(a) What is V_{out}?

(b) If $R = 1\,k\Omega$, what is the smallest R_L that the circuit can handle without "sagging" the output voltage?

(c) What should be the power rating of the resistor R? Be explicit about your assumptions.

Problem 3.2

Consider the half-wave rectifier circuit shown below. The ac input voltage to the circuit is a 60 Hz signal from a power transformer.



(a) What is the peak output voltage across the load resistor, assuming $R_L = 10\,\Omega$? Account for the voltage drop across the diode (look it up!).

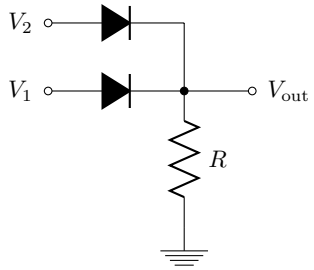
(b) Estimate the voltage ripple, assuming the same load resistance.

Problem 3.3

Sketch the analogous full-wave-rectifier-bridge circuit to the circuit in Problem 2, and repeat the calculations.

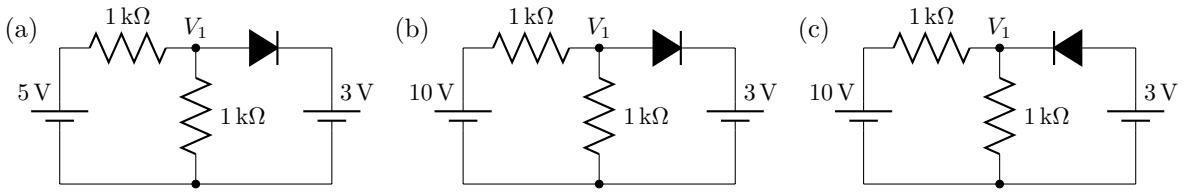
Problem 3.4

Consider the circuit below. Give the output voltage V_{out} in terms of the input voltages V_1 and V_2. You can assume ideal diodes (no forward voltage, breakdown, or leakage).



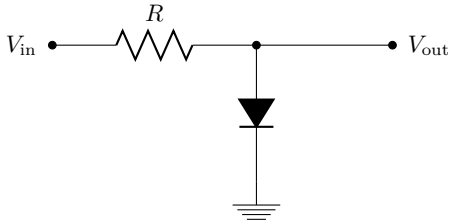
Problem 3.5

In each circuit, compute V_1 (relative to the negative battery terminals). Account for the forward-voltage drop across the diode as needed.



Problem 3.6

(a) Consider the "voltage-divider" circuit below, consisting of a resistor and a **forward-biased** (signal) diode.



Assuming $V_{\text{in}} > 0$, **use the diode law** to show that

$$V_{\text{in}} + I_{\text{S}}R = V_{\text{out}} + I_{\text{S}}R\,e^{V_{\text{out}}/nV_{\text{T}}}, \tag{3.3}$$

which determines V_{out} in terms of V_{in}, R, and T.

(b) Does the idea of a Thévenin-equivalent circuit make sense for this circuit? If yes, give the Thévenin-equivalent voltage and resistance. If no, explain *briefly* why not.

(c) Does V_{out} increase or decrease with T? (Assume R, I_{S}, and n are independent of temperature.)

(d) Obtain an explicit solution to the relation you derived in (a) as follows. Rearrange the equation to combine the parts that depend on V_{out} in the form $x\,e^x$, where

$$x = \frac{V_{\text{in}} - V_{\text{out}} + I_{\text{S}}R}{nV_T}. \tag{3.4}$$

That is, you should obtain an equation of the form $x\,e^x = y$, where y is independent of V_{out}. Then use the fact that the **Lambert W function** is defined by the relation

$$W(z)\,e^{W(z)} = z. \tag{3.5}$$

Thus, you should show that

$$V_{\text{out}} = V_{\text{in}} + I_{\text{S}}R - (nV_T)\,W\!\left(\frac{I_{\text{S}}R}{nV_T}\,e^{(V_{\text{in}}+I_{\text{S}}R)/nV_{\text{T}}}\right). \tag{3.6}$$

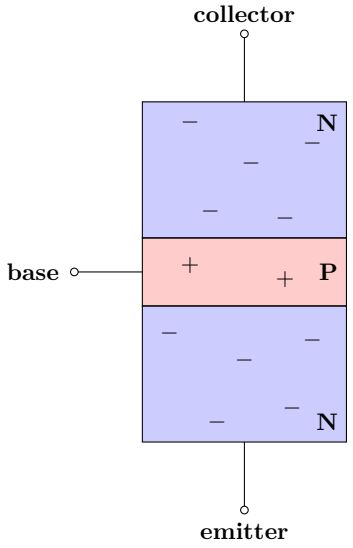
Chapter 4

Bipolar Junction Transistors

4.1 Overview

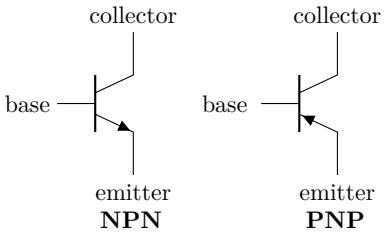
A **bipolar junction transistor (BJT)** is our first example of a device that is both *nonlinear* and *active*—active, in the sense that the device should be "powered," or to say it another way, it uses one signal to modify another signal. At first, it's a bit counterintuitive to have a device with *three* terminals, but roughly speaking, you can think of it functionally as having a pair of input terminals and a pair of output terminals, but one terminal is "shared" between the input and the output.

BJTs come in two flavors: NPN and PNP, which refers to the stack of doped semiconductors that form the transistor. The schematic construction of the NPN transistor is shown below: the name just gives the order of the layers from top to bottom (or bottom to top).



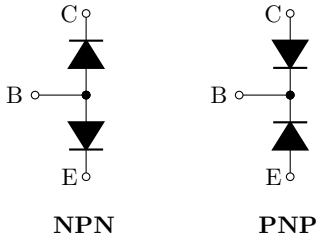
That is, the NPN transistor is a p-type semiconductor sandwiched in between two n-type semiconductors. However, the important thing is that the inner p-type layer is **thin** and **lightly doped**. The light doping means that there are relatively few p-type carriers in the p-type layer. The PNP transistor is pretty much the same thing, except for interchanging p-type and n-type layers. All the analysis here will go through to that case under this interchange, provided we also reverse all the currents and voltage differences. Thus we'll stick to the NPN case here (which is the more common case; usually, PNPs are only found when they are paired in some way with an NPN, in part because NPNs are easier to make well).

The schematic symbols for NPN and PNP transistors are shown below.



Note that the collector is distinguished from the emitter by the arrow, and the direction of the arrow distinguishes NPN from PNP transistors.

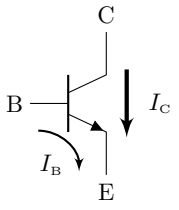
Looking at the BJT as a stack of alternately doped semiconductor layers, we can see that BJTs each have two p-n junctions, and we might expect these junctions to act like diodes. That is, you could reasonably expect BJTs to act like two diodes tied together, as shown below.



And, in fact, if you use a multimeter to measure continuity between the various leads, you will see continuity and voltage drops consistent with this simple "diode model." (This is in fact a good starting point for diagnosing a transistor that may be past its prime.) However, the difference is that the middle layer is a *single* region, and its thinness and low carrier density make it behave quite differently than a Siamese-twin pair of diodes.

4.2 Usage

The normal *modus operandi* of the (NPN) transistor is as shown below (again, the currents and voltage differences are reversed for the PNP).



Note the following:

1. All current goes out the emitter.
2. The base–emitter current I_B **controls** the collector–emitter current I_C.
3. For the currents to flow in the intended directions, $V_B > V_E$ and $V_C > V_E$.
4. The B–E junction acts like a *forward*-biased diode, and conducts current like a diode would (the arrow in the transistor symbol looks like a diode).
5. The C–E junction involves a *reverse*-biased diode. It *blocks* current, unless the B–E current overrides the blocking behavior.

6. For any device, there are limits to how large I_B, I_C, and V_{CE} can be, and these limits are different for each species of transistor.

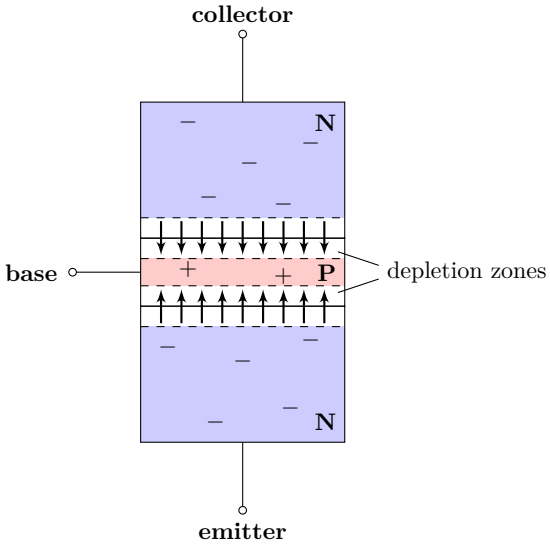
7. Under all the conditions above, the two currents are *proportional*:

$$I_C = \beta I_B. \tag{4.1}$$
(transistor current-control relation)

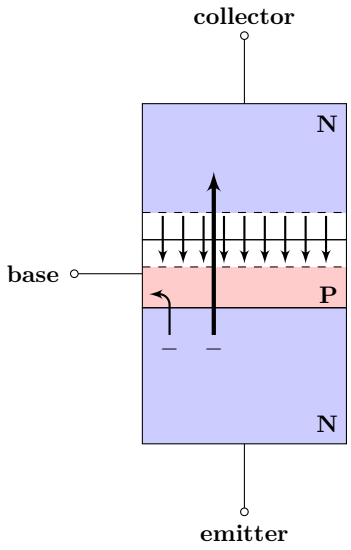
The parameter β is roughly constant, and as a simple value for estimating what happens in transistor circuits, you can assume $\beta \sim 100$. This parameter varies among different transistor species and even among individuals of one species. On transistor data sheets, β is often denoted by h_{BE}.

4.3 Mechanism

To understand why the transistor works the way it does, let's go back to the diagram of the stack of n- and p-type materials in the NPN transistor. The two p-n junctions set up two depletion zones, with corresponding electric fields as shown below.



The two depletion zones block any C–E current from flowing (in either direction). Then with voltages set up as $V_{BE} > 0.6\,\text{V}$ and $V_{CE} > 0$ (actually V_{CE} needs to be at least roughly $0.2\,\text{V}$), then the B–E depletion zone disappears, and the C–E depletion zone grows a bit.



Then, basically two things happen with the carriers, as in the diagram above.

1. Since current is flowing from base to emitter, there are n-type carriers flowing from the *emitter* to the *base* (the negative charge means they are flowing *against* the current). Remember the p-type region is lightly doped, so it is mostly n-type carriers that are transporting the current.

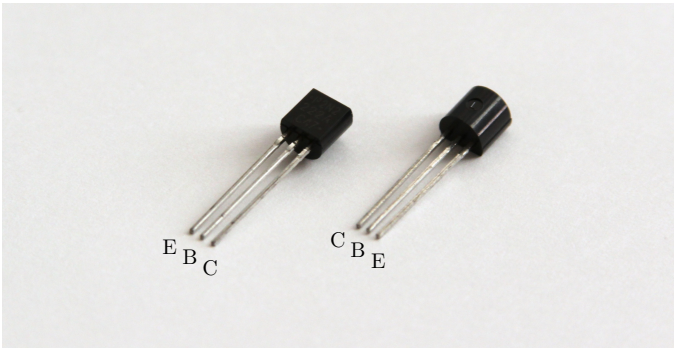
2. Many, or possibly most of, the n-type carriers never make it to the base terminal. What happens is that, once they are pulled into the p-type region by V_{BE}, they can diffuse into the C–B depletion zone, in which case the electric field in the depletion zone sweeps them into the collector's n-type region, leading to $I_{\text{C}} > 0$ (provided $V_{\text{CE}} > 0$, so the swept-up n-type carriers are removed through the collector terminal).

The transistor is a really beautiful device.

Note that from our description and diagrams, it would seem that the emitter and collector are basically equivalent, and it is true that they are very similar. However, in a real transistor, the geometry is very different, and the emitter material is heavily doped compared to the collector material (since it "produces" the carriers needed to transport the current). So while it is *possible* to operate a transistor with emitter and collector interchanged, it would not work nearly as well (the β in this configuration would be much smaller, for example).

4.4 Packaging

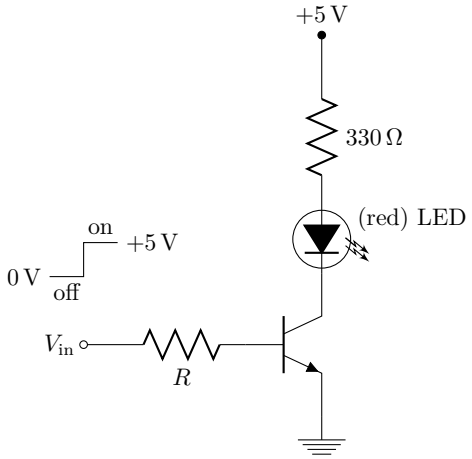
Transistors come in many shapes and sizes. The common TO-92 plastic case (appropriate for low-power signal transistors) is shown in the photograph below, with typical connections.



Note that these connections are common with 2N-series transistors in TO-92 packages, but they are *not* universal.[1]

4.5 Transistor Switch

Below is the first example circuit we'll do with transistors: using a transistor as a switch, here to turn a **light-emitting diode (LED)** on or off.



This switch works as advertised (for the stated input voltages) for a resistance R around 1 kΩ. To learn more about transistor operation, we will consider this resistance and a larger value for the base resistance to see two different regimes of transistor operation. This circuit behaves the same for any of a number of small-signal NPN transistors (2N2222A, 2N3904, 2N4401, etc.).

4.5.1 Saturation Mode

First, let's consider the case where $R = 1\,\text{k}\Omega$, which would be a typical way to design this switch. The "off" case is pretty easy: with 0 V input, $I_\text{B} = 0$, and thus $I_\text{C} = 0$ from Eq. (4.1), so the LED is off.

Now for the "on" case, with 5,V input. To start, the B–E junction of the transistor acts like a diode. Usually, the base current will be low, because it will switch a current β times larger, from Eq. (4.1). So we will assume diode drop across the B–E junction, but on the low end, say $V_\text{BE} = 0.6\,\text{V}$. Then the 1-kΩ resistor drops the rest of the input voltage, or 4.4 V. This gives a base–emitter current

$$I_\text{B} = \frac{4.4\,V}{1\,\text{k}\Omega} = 4.4\,\text{mA}, \tag{4.2}$$

which will control the collector current I_C.

Before considering the transistor action, let's consider the LED. LEDs have a higher forward voltage than signal diodes. Typical values for "bright" operation of a "T-1¾" sized LED (the most common size, with a plastic bulb 5 mm in diameter) depend on color:

- For red, orange, yellow, green-yellow: forward current $I_\text{F} = 20\,\text{mA}$; forward voltage $V_\text{F} = 1.8\,\text{V}$.
- For green, blue, white, UV: $I_\text{F} = 20\,\text{mA}$; $V_\text{F} = 3.3\,\text{V}$.

"High-brightness" LEDs can have larger forward currents, but these values are good for "normal" LEDs. We have a red LED, so $V_\text{F} = 1.8\,\text{V}$. If we think of the transistor as a switch that connects C–E (this is not quite true, more on this shortly), then the resistor drops $5 - 1.8 = 3.2\,\text{V}$; with a 330-Ω resistor, this gives I_F as a bit under 10 mA, which is reasonable, but hardly pushing the LED's brightness.

[1] In fact, the specific transistors in the photo should be wired *backwards* from this labeling. These are P2N2222A transistors by ON Semiconductor, and for some reason they decided to wire these CBE as viewed from the front, despite all *other* 2N2222A's in this package being wired EBC.

But what does the transistor *want* to do? Eq. (4.1) says that I_C should be β times I_B; assuming $\beta \sim 100$, then $I_C \sim 440\,\text{mA}$. But, as we found out, with the LED and resistor voltage drops, I_C can't possibly be this high (the transistor can't have a *negative* voltage drop, for example). In fact, the transistor drop V_{CE} can't be less than about 0.2 V, so the resistor actually drops about 3.0 V. With the 330-Ω resistor, this gives $I_C = I_F \approx 9.1\,\text{mA}$.

This mode of transistor operation is called **saturation mode**: as a switch, the transistor is "wide open," and the current I_C is limited by external elements (here, LED and 330-Ω resistor), not the transistor. Of course, the current-limiting resistor could be reduced somewhat (say, to 150 Ω) for a brighter LED here.

The important thing to notice in this example is that a high-impedance source (1 kΩ, from the base resistor, plus any impedance of the input voltage source) controls a higher-current load via the transistor.

4.5.2 Forward-Active Mode

Now suppose $R = 100\,\text{k}\Omega$ in the same circuit. Then the base current is two orders of magnitude smaller, or $44\,\mu\text{A}$. The transistor tries to set $I_C \sim 100 I_B = 4.4\,\text{mA}$, and now this is certainly possible for the transistor. Just to double check, the voltage drop across the 330-Ω resistor is $330\,\Omega \times 4.4\,\text{mA} = 1.5\,\text{V}$. Then removing the resistor and LED drops, $V_{CE} = 5\,\text{V} - 1.8\,\text{V} - 1.5\,\text{V} = 1.7\,\text{V}$. Now it is the *transistor* that is regulating the current via its voltage drop. This is called the **forward-active mode** of the transistor, and it is in this regime that the transistor can act as an amplifier with some gain (the saturation mode is a kind of "infinite-gain" amplification).

4.5.3 Summary

Let's just summarize the difference between saturation and forward-active modes in the switching circuit, because this can be a bit confusing the first time through, and it's important to understand the difference *intuitively*.

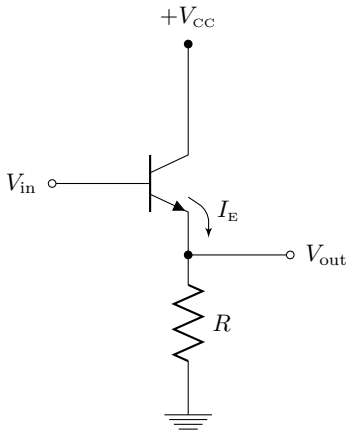
In **forward-active mode**, the base current I_B controlled the collector current I_C (and thus the LED current) via the relation $I_C = \beta I_B$. The transistor "regulates" I_C by adjusting the voltage V_{CE} to the proper amount. We know the current is set by the voltage drop across the 330-Ω resistor: the larger the resistor drop, the larger the current. This means that V_{CE} is *smaller* for larger currents, because more of the supply voltage must be taken up by the resistor.

In **saturation mode**, the transistor relation *doesn't* hold. That's because the base current makes the transistor "want" more current than the 330-Ω resistor will allow. The transistor normally tries to drive more current by reducing V_{CE}, but it can't do so below zero (or more precisely, below about 0.2 V if $I_C > 0$), so $I_C < \beta I_B$.

So you can really think of the transistor relation as being more like $I_C \leq \beta I_B$. Both the transistor and an external element (resistor) will want to limit the current; whichever wants *less* current is the one that wins.

4.6 Emitter Follower

As our second transistor circuit, we'll continue with a transistor in forward-active mode. Below is the **emitter follower**, where the output voltage is intended to "follow," or match, the input voltage. It is an amplifier in the sense of being an amplifier for *current*.



Note that we are introducing some new notation here. In addition to the base and collector currents in the diagram on p. 72, we are introducing the emitter current I_E. Also, the power-supply voltage is labelled $+V_{CC}$; the plus denotes a positive voltage with respect to ground and the "C" subscript denotes this is intended to power the collector terminal (two C's distinguishes this from the voltage V_C at the collector, which is the same in this circuit, but not always).

We will consider the signals to be *ac* signals, with some dc bias that we don't care much about. This is kind of a fact of life with transistors, which can only work with current flowing in one direction—if we want to handle a bipolar signal (like an audio signal), these biases ensure that the net signal is compatible with the way the transistor works. The bias voltage/current is typically rejected at some point near the output of the circuit, for example using a high-pass filter.

Thus, let's set

$$V(t) = V_0 + v(t), \qquad I(t) = I_0 + i(t), \tag{4.3}$$
(dc and ac components)

where V_0 and I_0 are the dc biases, and the (small) ac signals are $v(t)$ and $i(t)$. This simplifies the analysis a bit, because for example we can write the output voltages as

$$V_{out} = V_E = V_B - 0.6\,\mathrm{V}, \tag{4.4}$$

but dropping the dc offsets, this becomes

$$v_{out} = v_E = v_B. \tag{4.5}$$

The **ac voltage gain** for the circuit is defined by

$$G := \frac{v_{out}}{v_{in}} = \frac{v_E}{v_B} = 1, \tag{4.6}$$
(ac voltage gain)

so this is indeed a unity-gain circuit (voltage follower).

To compute the current gain, we can relate the emitter current to the base and collector currents by

$$I_E = I_B + I_C = I_B + \beta I_B = I_B(\beta + 1), \tag{4.7}$$

where we used Eq. (4.1). In terms of ac components, this is

$$i_E = i_B(\beta + 1), \tag{4.8}$$

so the **ac current gain** is $(\beta + 1)$.

4.6.1 Input and Output Impedance

Since the emitter follower has current gain, it can allow a source with high output impedance to drive a lower-impedance load. To quantify this, let's define Z_{load} to be the impedance at the transistor emitter,

which is R in parallel with any other load impedance attached to the V_{out} terminal. Then we can calculate

$$i_{\text{B}} = \frac{i_{\text{E}}}{\beta+1} = \frac{v_{\text{E}}}{Z_{\text{load}}(\beta+1)} = \frac{v_{\text{B}}}{Z_{\text{load}}(\beta+1)}. \tag{4.9}$$

Here, we used the current gain (4.8), then $v_{\text{E}} = i_{\text{E}}R$, then $v_{\text{E}} = v_{\text{B}}$. Then we can define the **input impedance** of the amplifier by

$$Z_{\text{in}} := \frac{v_{\text{in}}}{i_{\text{in}}}. \tag{4.10}$$
(input impedance, definition)

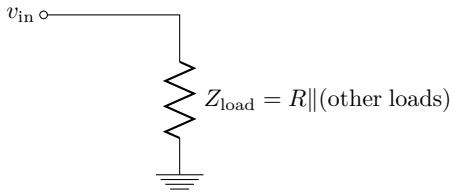
Writing the input voltage and current in terms of the base voltage and current,

$$Z_{\text{in}} = \frac{v_{\text{B}}}{i_{\text{B}}}, \tag{4.11}$$

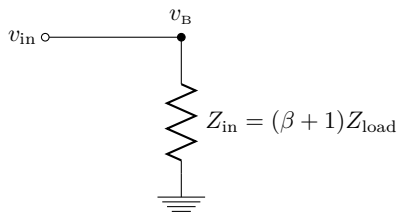
and then using Eq. (4.9),

$$Z_{\text{in}} = Z_{\text{load}}(\beta+1). \tag{4.12}$$
(input impedance, emitter follower)

We can interpret this as follows: *Without* the transistor, the input v_{in} would have to "drive" the load impedance (the resistor R and other connected stuff) directly, as shown below



The effect of the transistor is to multiply the resistor value by $(\beta+1)$, so effectively the impedance is about 100 times larger, and thus much easier to drive—the transistor is supplying most of the current via the collector, so the voltage source driving the resistor doesn't have to supply much current at all. Thus, the equivalent circuit *with* the transistor is shown below.



Of course, we can think of this (Thévenin) equivalent circuit because the transistor is acting as a linear amplifier.

We can also define an **output impedance** as

$$Z_{\text{out}} := \frac{v_{\text{out}}}{i_{\text{out}}}. \tag{4.13}$$
(output impedance, definition)

Manipulating this a bit,

$$Z_{\text{out}} = \frac{v_{\text{in}}}{i_{\text{E}}} = \frac{v_{\text{in}}}{i_{\text{B}}(\beta+1)}, \tag{4.14}$$

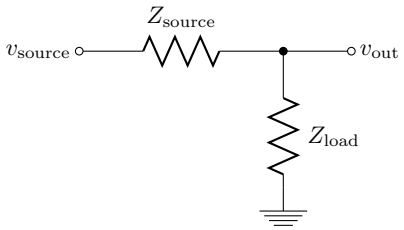
and then defining the source impedance

$$Z_{\text{source}} := \frac{v_{\text{in}}}{i_{\text{B}}} \tag{4.15}$$
(impedance of input source)

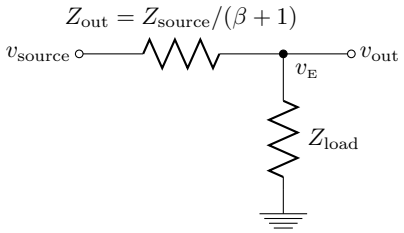
(i.e., the impedance of whatever is driving the transistor base at voltage V_{in}), the output impedance becomes

$$Z_{\text{out}} = \frac{Z_{\text{source}}}{\beta+1}. \tag{4.16}$$
(output impedance, emitter follower)

We can interpret this as follows: *Without* the transistor in place, the load would be connected directly to the source, loading it down somewhat:



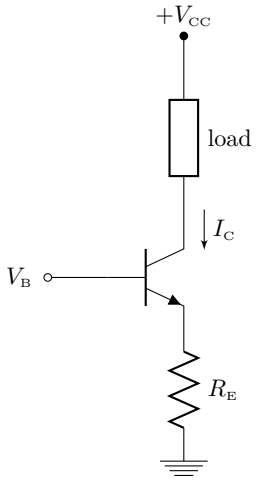
That is, v_{source} and Z_{source} represent the Thévenin-equivalent voltage and impedance of the signal source. The transistor effectively divides the source impedance by $(\beta + 1)$, making the source effectively "stiffer" (closer to an ideal voltage source) by about 100-fold, as "seen" by the load. Equivalently, the transistor is supplying most of the load current, so the source only has to supply a small fraction of the current that it otherwise would.



Again, from this point of view, the transistor is "lumped into" the source circuit, which is appropriate because it is acting as a linear device.

4.7 Transistor Current Source

The next circuit we will consider is the **transistor current source**. This is more of a prelude to the transistor amplifier that is coming up next, but this is also useful in its own right. A **current** source is a circuit that maintains a constant (programmed) current, independent of voltage, within limits of course. In the circuit below, the goal is to maintain a constant current through the load, independent of the load impedance.



In this case, the transistor is maintaining the collector current I_{C}. To see that this works, note that the base and emitter voltages are related as usual by

$$V_{\text{E}} = V_{\text{B}} - 0.6\,\text{V}. \tag{4.17}$$

Then the emitter resistor R_E sets the emitter current via

$$I_E = \frac{V_E}{R_E} = \frac{V_B - 0.6\,\text{V}}{R_E}. \tag{4.18}$$

Remember that the base and collector currents add to form the emitter current. Since $I_C = \beta I_B$, then

$$I_E = I_C + I_B = I_C(1 + \beta^{-1}) \approx I_C, \tag{4.19}$$

provided β is large. Thus,

$$I_C \approx \frac{V_B - 0.6\,\text{V}}{R_E}. \tag{4.20}$$
(transistor current source)

That is, the load current I_C is programmed by the input base voltage V_B, as well as the emitter resistor R_E. Importantly, the question of the load resistance never came into the analysis, so the current is independent of the load resistance.

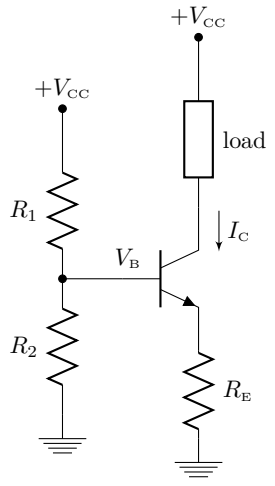
4.7.1 Compliance

Well, that is, within limits. What are the limits? That is, what is the **compliance** of the current source? Specifically, the compliance refers to the range of output voltages for which the transistor is properly regulating the current—the output voltage here meaning the collector voltage V_C, which the transistor "presents" to the load.

For proper transistor operation, we need $V_B > V_E$ to switch the collector current, and we also need the collector $V_C \gtrsim V_E + 0.2\,\text{V}$. On the upper end (i.e., lower-current end), we can have V_C going all the way up to $+V_{CC}$, assuming this doesn't exceed the breakdown voltage of the transistor. So, for a given input voltage V_B, the range of output voltages V_C is from $V_E + 0.2\,\text{V} = (V_B - 0.6\,\text{V}) + 0.2\,\text{V} = V_B - 0.4\,\text{V}$, on up to $+V_{CC}$. This corresponds to a range of current from 0 on up to $(V_{CC} - V_B + 0.4\,\text{V})/R_{\text{load}}$, in terms of the load resistance.

4.7.2 Bias Network

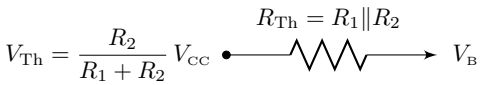
Of course, to make the current source *work*, we need to set the base voltage V_B. How do we do this if we only have ground and the power-supply voltage? The answer, of course, is a voltage divider, as in the circuit below.



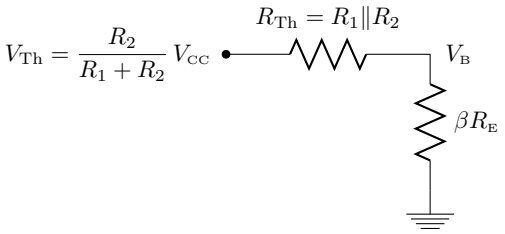
The voltage divider here could even be a potentiometer (variable resistor), so we can fine-tune the regulated current (our treatment is only approximate, for example, because we are using the approximate V_{BE} drop of 0.6 V). The voltage divider gives an *unloaded* voltage of

$$V_B \approx \frac{R_2}{R_1 + R_2} V_{CC}, \tag{4.21}$$

but we are "loading" the divider with the transistor, so we have to be careful. Remember that the Thévenin equivalent circuit for the voltage divider (Section 1.4.1) is as follows:



In our discussion of the emitter follower, we saw that the input impedance of the transistor is about βR_{E}. That is, as "seen" from the base-side of the transistor, the equivalent input circuit is



Here, we have the input voltage divider, still connected to the transistor base, but with the transistor replaced by its effective input impedance. Thus, V_{B} is determined by another voltage divider, but V_{B} is essentially the unloaded value above (i.e., V_{Th}) under the condition

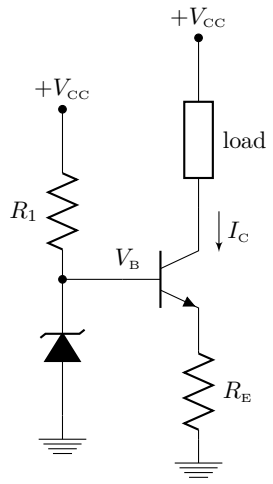
$$R_{\text{Th}} = R_1 \| R_2 \ll \beta R_{\text{E}}. \qquad (4.22)$$

(design condition for bias network)

This condition ensures that the voltage divider is "stiff"—that is, it acts like an ideal voltage source. If this condition is not fulfilled, the divider's voltage "sags" under the load of R_{E} via the transistor (i.e., the input impedance of the transistor).

Note that in designing this circuit, you might think that we can just *calculate* the effect of R_{E} on V_{B}, and so we can just tweak the ratio of R_1 to R_2 to compensate. However, the resulting voltage depends on β, which can vary from device to device, or with temperature, or with collector current, etc. That is, in this circuit design (and other circuit designs), it's important to be in the regime where we can neglect the effect of R_{E} on V_{B}—that is, in the regime where β is large, but the *particular value* of β is not critical.

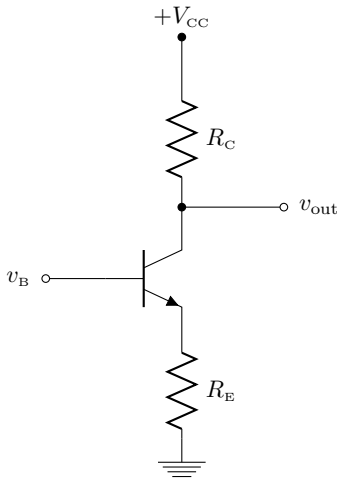
One other solution to the problem of a sagging voltage divider is to use the Zener-diode voltage regulator (Section 3.5) to set V_{B}.



In this case, we can get a fairly coarse choice of V_{B} by choosing the Zener diode, and fine-tune the current by setting R_{E}.

4.8 Common-Emitter Amplifier

And now for our first voltage amplifier that has **gain**—the output can be larger than the input. The basic circuit is simple: just take a transistor current source, and use a collector resistance R_C as the load. Heuristically, the input voltage programs a collector current I_C, and R_C acts to convert the current back into a voltage, which serves as the output.



This amplifier only works as advertised on ac signals, so let's consider small ac variations $v(t)$ and $i(t)$ with respect to dc biases V_0 and I_0, as before:

$$V(t) = V_0 + v(t), \qquad I(t) = I_0 + i(t). \tag{4.23}$$

The current-source result (4.20), dropping dc biases, becomes

$$i_C = \frac{v_B}{R_E}. \tag{4.24}$$

The voltage drop across R_C is

$$V_{CC} - V_C = I_C R_C, \tag{4.25}$$

which, in terms of ac quantities, is

$$v_C = -i_C R_C. \tag{4.26}$$

Combining this with Eq. (4.24),

$$v_C = -\frac{R_C}{R_E}\, v_B. \tag{4.27}$$

Identifying these voltages with the input and output voltages,

$$v_{out} = -\frac{R_C}{R_E}\, v_{in}. \tag{4.28}$$
(common-emitter amplifier)

Defining the **gain** of the amplifier by

$$G := \frac{v_{out}}{v_{in}}, \tag{4.29}$$
(ac amplifier gain)

the common-emitter gain is

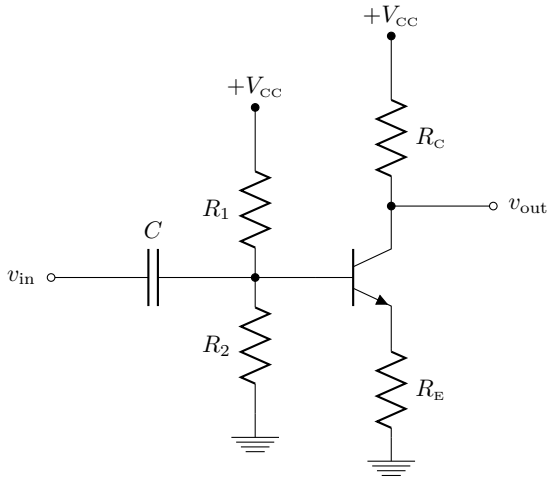
$$G = -\frac{R_C}{R_E}. \tag{4.30}$$
(common-emitter gain)

Then we see that the gain is *negative*, meaning the input ac signal is inverted at the output, and the ratio of collector and emitter resistors controls the gain.

For the sample-circuit numbers, we have $G = -10$.

4.9 Bias Network (AC Coupling)

As in the current source, we need an input network to set the dc bias. A nonzero bias is critical for the amplification of the ac signal: the B–E junction can only conduct in the forward direction, and so the input signal can't cross through zero and still be amplified without a lot of distortion.



Here, the voltage divider sets the bias voltage, and the input capacitor only passes the ac part of the input signal.

Let's go through the different parts of the circuit, and work out all the relevant parameters. As a concrete example, we will use the parameters:

$$\begin{aligned} V_{CC} &= 15\,\text{V}, \\ R_1 &= 56\,\text{k}\Omega, \\ R_2 &= 5.6\,\text{k}\Omega, \\ C &= 0.1\,\mu\text{F}, \\ R_E &= 330\,\Omega, \\ R_C &= 3.3\,\text{k}\Omega. \end{aligned} \tag{4.31}$$

1. **AC input impedance.** The input circuit is a high-pass filter, with capacitance C. The Thévenin resistance is the parallel resistance of R_1 and R_2, but we must also include the input impedance βR_E of the amplifier as an additional parallel resistance. Thus, at high frequencies when the capacitor acts as a short-circuit, the input impedance of the circuit is the Thévenin resistance, or $R_1 \| R_2 \| \beta R_E$.

   For this circuit, $R_1 \| R_2 = 5.1\,\text{k}\Omega$. Also, $\beta R_E \approx 100\,\text{k}\Omega$, which is much larger than $R_1 \| R_2$, so $R_1 \| R_2 \| \beta R_E \approx R_1 \| R_2 = 5.1\,\text{k}\Omega$.

2. **High-pass input.** The corner frequency of the input network is

   $$f_{3\text{dB}} = \frac{1}{2\pi (R_1 \| R_2 \| \beta R_E) C}. \tag{4.32}$$
   (input corner frequency)

   For the circuit here, $f_{3\text{dB}} = 310\,\text{Hz}$.

3. **Loading condition.** It is good practice for the input impedance of the transistor to have negligible effect. That is, we should have

   $$\beta R_E \gg R_1 \| R_2, \tag{4.33}$$

   in which case the input impedance and corner frequency of the input network do not depend on β (and thus on temperature, etc.).

   For the sample numbers here, we have already seen that we satisfy this condition.

4. **Gain.** From Eq. (4.30), the gain of the amplifier is

$$G = -\frac{R_C}{R_E}. \tag{4.34}$$

5. **Output impedance.** Because the transistor acts like a current source (with I insensitive to V), the transistor effectively presents a very large impedance at the output—a large change in voltage causes *little* change in the current. Thus the impedance at the output is the Thévenin resistance at that point, or the parallel impedance of the the collector with R_C. Thus, we can take the output impedance of the amplifier to be approximately R_C.

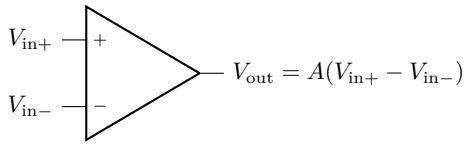
   In this example, the output impedance of the amplifier is $3.3\,\mathrm{k}\Omega$.

6. **Bias points.** The bias at the input is just the voltage-divider voltage, including any correction from βR_E, which we usually want to ignore. In the circuit here, the 15 V supply is divided down to 1.36 V.

   Why does this make sense? It's important to note that the bias at the input isn't critical in the sense of needing to be near the middle of the supply range. If we are doing a lot of amplification and want to avoid **clipping**, though, the output should be close to the center of the supply. Let's see if this is the case. The emitter voltage is $V_E = V_B - 0.6\,\mathrm{V} = 0.76\,\mathrm{V}$. This programs a collector current $I_C \approx (V_B - 0.6\,\mathrm{V})/R_E = 2.3\,\mathrm{mA}$, from Eq. (4.20). Then the voltage drop across R_C is 7.6 V which is about half of V_{CC}. Notice how the matching ratios of R_1 to R_2 and R_C to R_E lead to a sensible bias voltage here, but because the (fixed) voltage V_{BE} enters here, this rule does not always apply.

4.10 Transistor Differential Amplifier

Now we come to a more sophisticated transistor amplifier, the **transistor differential amplifier**. Schematically, a differential amplifier has the following form:



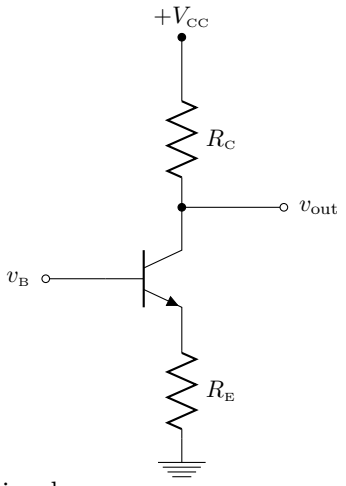
That is, the output is the difference of the inputs (the signs at the inputs tell you which is subtracted from which), and multiplied by the voltage **gain factor** A.

Why is a differential amplifier useful? For example:

1. In the transmission of signals, differential amplifiers give you **noise immunity**— if you transmit a signal on two lines (signal and ground), the same noise appears on each line, and gets cancelled out by a differential amplifier on the receiving end.

2. Differential amplifiers offer a straightforward way to implement **negative feedback**, which is a nice way to achieve close-to-ideal behavior in nearly every respect. We will go over this in much more detail when we get to op-amps.

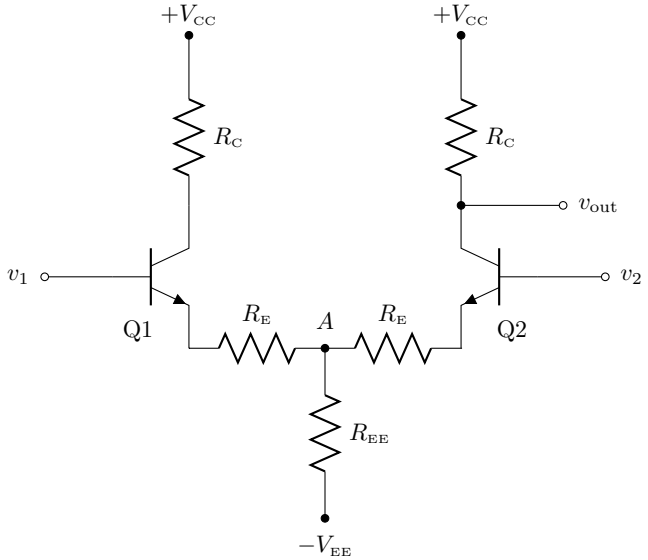
The simplest realization of a differential amplifier uses two transistors. To understand this, first recall the common-emitter amplifier.



In this amplifier, the result for the ac signal was

$$v_{out} = -\frac{R_C}{R_E} v_{in}. \tag{4.35}$$

The differential amplifier is basically a "stack" of two common-emitter amps, sharing a common resistance R_{EE} at the negative end.



From the layout of the schematic, you can see why this circuit is also called a **long-tailed pair** (the "long" here refers to the magnitude of R_{EE}, which as we will see is typically large compared to R_E). Note also that the emitter-end of the circuit is powered by a negative supply, so we can have positive or negative outputs.

To analyze this circuit, we will again concentrate on the ac components of the inputs, v_1 and v_2. Now let's think of v_1 and v_2 as being deviations from the mean voltage $\bar{v}$:

$$v_1 = \bar{v} + \frac{\Delta v}{2}, \qquad v_2 = \bar{v} - \frac{\Delta v}{2}. \tag{4.36}$$

Here, $\bar{v}$ is the **common-mode signal**,

$$\bar{v} := \frac{v_1 + v_2}{2}, \tag{4.37}$$

and δv is the **differential signal**,

$$\Delta v := v_1 - v_2. \tag{4.38}$$

Ideally, a differential amplifier responds *only* to the differential signal, and is insensitive to the common-mode signal.

4.10.1 Differential-Only Input

We will show that the circuit responds linearly to the inputs, so we can treat the differential and common-mode signals *separately*, just add them together to handle an *arbitrary* input signal. So first, let's concentrate on only the differential signal. That is, setting $\bar{v} = 0$, we have

$$v_1 = \frac{\Delta v}{2}, \qquad v_2 = -\frac{\Delta v}{2}. \tag{4.39}$$

Then, just as in the emitter follower,

$$v_{\text{E}1} = v_{\text{B}1} = v_1, \tag{4.40}$$

and similarly

$$v_{\text{E}2} = v_2. \tag{4.41}$$

At point A in the circuit, ignoring dc offsets, we have a 50% voltage divider between $v_{\text{E}1}$ and $v_{\text{E}2}$, so the voltage is the average of these:

$$v_A = \frac{v_{\text{E}1} + v_{\text{E}2}}{2} = \frac{v_1 + v_2}{2} = 0. \tag{4.42}$$

We can interpret this as follows: The point A acts like the "ground" point for the two common-emitter amplifiers in this circuit. Since this point is stable with respect to differential inputs, the common-emitter results carry through here, and in particular for the right-hand common-emitter amp,

$$v_{\text{out}} = -\frac{R_{\text{C}}}{R_{\text{E}}} v_2, \tag{4.43}$$

or in terms of the differential signal,

$$v_{\text{out}} = \frac{R_{\text{C}}}{2R_{\text{E}}} \Delta v. \tag{4.44}$$
(differential response)

Note that the minus sign disappeared here, which is why it is sensible to take the output from the right-hand transistor. Thus, we have

$$G_{\text{diff}} := -\frac{R_{\text{C}}}{2R_{\text{E}}}. \tag{4.45}$$
(differential gain factor)

as the **differential gain factor** for the amplifier.

4.10.2 Common-Mode-Only Input

Now we can focus on *just* the common-mode signal. That is, we take $\Delta v = 0$, so that

$$v_1 = v_2 = \bar{v}. \tag{4.46}$$

Then applying Kirchoff's law to the currents at point A,

$$i_{\text{EE}} = i_{\text{E}1} + i_{\text{E}2} =: 2i_{\text{E}}, \tag{4.47}$$

since the two emitter currents are the same. Here, $i_{\text{E}1}$ and $i_{\text{E}2}$ are the (ac components of the) currents out each emitter (both equal to i_{E}), and i_{EE} is the current through R_{EE}. Then applying Ohm's law at point A,

$$v_A = i_{\text{EE}} R_{\text{EE}} = 2i_{\text{E}} R_{\text{EE}}, \tag{4.48}$$

and now Ohm's law across either emitter resistor gives

$$i_{\text{E}} = \frac{v_{\text{E}} - v_A}{R_{\text{E}}} = \frac{v_{\text{in}} - 2i_{\text{E}} R_{\text{EE}}}{R_{\text{E}}}, \tag{4.49}$$

after eliminating v_A. Solving for i_{E},

$$i_{\text{E}} = \frac{v_{\text{in}}}{R_{\text{E}} + 2R_{\text{EE}}}. \tag{4.50}$$

Then the output is, just as we had in the common-emitter amplifier,

$$v_{\text{out}} = -i_{\text{C}} R_{\text{C}} \approx -i_{\text{E}} R_{\text{C}}, \tag{4.51}$$

if we assume β large, and then using Eq. (4.50),

$$v_{\text{out}} = -\left(\frac{R_{\text{C}}}{R_{\text{E}} + 2R_{\text{EE}}}\right) v_{\text{in}}, \tag{4.52}$$
(common-mode response)

such that we can define

$$G_{\text{CM}} := -\frac{R_{\text{C}}}{R_{\text{E}} + 2R_{\text{EE}}} \tag{4.53}$$
(common-mode gain factor)

as the **common-mode gain factor**.

4.10.3 General Input and Common-Mode Rejection

Since the circuit responds linearly to the inputs, we can take a general pair of inputs v_1 and v_2, decompose them into differential and common-mode components via Eqs. (4.37) and (4.38), and then the output is

$$v_{\text{out}} = G_{\text{diff}} \Delta v + G_{\text{CM}} \bar{v}. \tag{4.54}$$
(common-mode gain factor)

Again, for a "good" differential amplifier, G_{CM} should be zero, or at least small compared to G_{diff}.

One way to quantify the "goodness" of the differential amplifier is the **common-mode rejection ratio (CMRR)**, which we define as

$$\text{CMRR} := \left|\frac{G_{\text{diff}}}{G_{\text{CM}}}\right| = \frac{R_{\text{E}} + 2R_{\text{EE}}}{2R_{\text{E}}}. \tag{4.55}$$
(common-mode rejection ratio)

Thus, typical differential-amplifier designs will be such that $R_{\text{EE}} \gg R_{\text{E}}$, in which case the CMRR reduces to the simple ratio

$$\text{CMRR} \approx \frac{R_{\text{EE}}}{R_{\text{E}}}. \tag{4.56}$$

Again, an ideal differential amplifier has a large CMRR. Typically the CMRR is large enough that it is usually measured in dB.

To give an intuitive recap, a *differential* input signal produces an output signal by changing the currents i_{E1} and i_{E2} through the emitter resistors R_{E}. But these currents cancel at point A, the current through R_{EE} stays fixed, and the two transistors act like common-emitter amplifiers. To maximize the differential gain, we want R_{E} to be much smaller than R_{C}. A *common-mode* signal, however, *does* change i_{EE}. However, if we make R_{EE} large, then the change in current i_{E} due to a common-mode signal is small. The change in the collector current i_{C} and thus output voltage is correspondingly small for a common-mode input, making the amplifier relatively insensitive to common-mode signals.

4.10.4 Improving the Differential Amplifier

One problem with the differential amplifier is that it is usually desirable to have a large differential gain G_{diff}, which means R_{C} should be large. However, R_{C} is also the output impedance, and we usually *don't* want a large output impedance (we want the output to act more like an ideal voltage source). A simple solution to this is to buffer the output using an emitter follower, which will reduce the effective output impedance by a factor of β.

Another observation is that R_{EE} should be large for a high CMRR. An improvement is to replace R_{EE} with a transistor current source. This regulates a constant i_{EE}, which regulates a constant i_{E} in the common-mode analysis. Since i_{E} is acting as a constant, this implies a very large effective R_{EE} in Eq. (4.50).

Finally, note that we can obtain a good CMRR by having $R_{\text{E}} \ll R_{\text{EE}}$. In fact, why can't we just set $R_{\text{E}} = 0$? We *can* replace the resistors R_{E} by shorts, but still these resistances will not be zero. In this case, the resistances will be the **intrinsic resistance** r_{e} of the transistors, which we will return to below.

4.11 Ebers–Moll Equation

So far, we have been understanding transistor circuits using the crude **current-control model**, centered on the equation [Eq. (4.1)]

$$I_{\mathrm{C}} = \beta I_{\mathrm{B}}, \tag{4.57}$$

with β roughly constant. The crudeness is, of course, that β is *not* constant. Usually we try to design in such a way that this doesn't matter, but it is still useful to develop a better model of the transistor.

Recall the diode law (3.1), relating the diode current and voltage:

$$I = I_{\mathrm{S}} \left(e^{V/nV_{\mathrm{T}}} - 1 \right). \tag{4.58}$$

Here, I_{S} is the saturation current (reverse leakage current), n is the ideality, V_{T} is the **thermal voltage**

$$V_{\mathrm{T}} := \frac{k_{\mathrm{B}} T}{e}, \tag{4.59}$$

$k_{\mathrm{B}} = 1.381 \times 10^{-23}\,\mathrm{J/K}$ is the Boltzmann constant, and $e = 1.602 \times 10^{-19}\,\mathrm{C}$ is the fundamental charge. The base–emitter junction of the transistor works much like a diode (recall that the collector current arises as a "side effect" of the base–emitter current, since carriers headed initially towards the base are swept into the collector), and so a similar law applies to transistor operation, the **Ebers–Moll equation**:

$$I_{\mathrm{C}} = I_{\mathrm{CS}} \left(e^{V_{\mathrm{BE}}/nV_{\mathrm{T}}} - 1 \right). \tag{4.60}$$
(Ebers–Moll equation)

This relates the collector current I_{C} to the base–emitter voltage $V_{\mathrm{BE}} = V_{\mathrm{B}} - V_{\mathrm{E}}$, via a saturation current I_{CS}. Unlike the current-control view of the β model of the transistor, the Ebers–Moll equation gives a **voltage-control view** of transistor operation. The control of I_{C} via I_{B} is only *indirect*, since I_{B} is controlled by V_{BE} via diode-like conduction. Thus, the transistor is a **transconductance device** (meaning a device that converts voltage to current). The Ebers–Moll model turns out to be accurate over a wide range of currents, typically from nA–mA. Note that in what follows, we will drop the ideality factor n to simplify notation somewhat without significantly affecting the calculations.

4.11.1 Magnitudes

To compare with what we know before, for a typical circuit analysis we assumed $V_{\mathrm{BE}} \approx 0.6\,\mathrm{V}$, which is much bigger than the thermal voltage V_{T} (25.69 mV at 25°C). Thus the exponential in the Ebers–Moll equation is large compared to the 1, and so Eq. (4.60) becomes

$$I_{\mathrm{C}} \approx I_{\mathrm{CS}} e^{V/V_{\mathrm{T}}} \tag{4.61}$$

and

$$I_{\mathrm{C}} \gg I_{\mathrm{CS}}. \tag{4.62}$$

This also means that, as in the diode, that V_{BE} is relatively insensitive to variations in the collector current, if we think about the relation in a "backwards" way. That is, if I_{C} changes by a factor of 10, and V_{BE} changes by ΔV_{BE}, then

$$10 \approx e^{\Delta V_{\mathrm{BE}}/V_{\mathrm{T}}}, \tag{4.63}$$

so that the voltage change is

$$\Delta V_{\mathrm{BE}} \approx V_{\mathrm{T}} \log 10, \tag{4.64}$$

which is about 59 mV at 25°.

4.11.2 Relation to β

Then how to we recover the current-control relation (4.57)? If we think of the base-emitter junction as a diode, then we simply apply the diode law (4.58) to obtain

$$I_{\mathrm{B}} = I_{\mathrm{S}} \left(e^{V_{\mathrm{BE}}/V_{\mathrm{T}}} - 1 \right), \tag{4.65}$$

where I_{S} is the saturation current of the base–emitter junction, and we have set $n \approx 1$. Then solving this for V_{BE},

$$V_{\mathrm{BE}} = V_{\mathrm{T}} \log \left(\frac{I_{\mathrm{B}}}{I_{\mathrm{S}}} + 1 \right), \tag{4.66}$$

and putting this into the Ebers–Moll equation, we find

$$I_{\mathrm{C}} = \left(\frac{I_{\mathrm{CS}}}{I_{\mathrm{S}}} \right) I_{\mathrm{B}}. \tag{4.67}$$

This has the form of the transistor β relation (4.57), with a β of $I_{\mathrm{S}}/I_{\mathrm{CS}}$. Note, however, that β is temperature dependent and somewhat dependent on current, which is not reflected in this simple derivation (for example, the proportionality is not exact if we do not set $n = 1$ in the diode law).

4.11.3 Intrinsic Emitter Resistance

The Ebers–Moll equation is also useful in establishing an **intrinsic emitter resistance** of the transistor. For small ac signals, the intrinsic resistance in the emitter is defined by

$$v_{\mathrm{BE}} = i_{\mathrm{E}} r_{\mathrm{e}}. \tag{4.68}$$

More generally, we can regard the resistance to be defined by the I–V slope

$$\frac{1}{r_{\mathrm{e}}} := \frac{dI_{\mathrm{E}}}{dV_{\mathrm{BE}}}, \tag{4.69}$$

(intrinsic emitter resistance: definition)

thinking of the current as a response to the voltage. That is, an ac emitter current i_{E} modulates the base-emitter voltage v_{BE} by an amount controlled by r_{e}. Given that these ac signals are small, this relation is given by the derivative of the v_{BE}–i_{E} relation, or the Ebers–Moll equation. We can calculate this via

$$\frac{1}{r_{\mathrm{e}}} = \frac{i_{\mathrm{E}}}{v_{\mathrm{BE}}} \approx \frac{i_{\mathrm{C}}}{v_{\mathrm{BE}}}, \tag{4.70}$$

since we assume β is large, as usual. Then at some bias collector current I_{C},

$$\frac{1}{r_{\mathrm{e}}} \approx \frac{dI_{\mathrm{C}}}{dV_{\mathrm{BE}}}. \tag{4.71}$$

Differentiating Eq. (4.61), this becomes

$$\frac{1}{r_{\mathrm{e}}} \approx \frac{I_{\mathrm{C}}}{V_{\mathrm{T}}}, \tag{4.72}$$

or

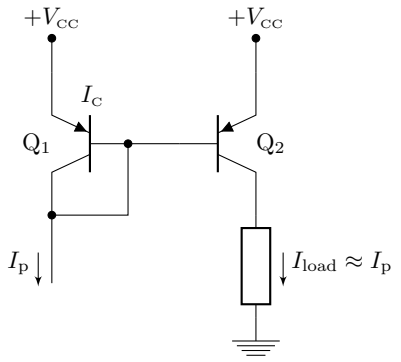
$$r_{\mathrm{e}} \approx \frac{V_{\mathrm{T}}}{I_{\mathrm{C}}} \approx \frac{25.69\,\mathrm{mV}}{I_{\mathrm{C}}} \text{ at } 20^{\circ}\mathrm{C}. \tag{4.73}$$

(intrinsic emitter resistance)

This gives a useful expression for the intrinsic resistance of the transistor, which depends on both current and temperature.

4.11.4 Current Mirror

A good example of a simple transistor circuit that can best be understood via the Ebers–Moll equation is the **current mirror**. Essentially, I_p in the circuit below is the "program current," and the mirror "copies" the current through the load, independent of the load impedance (within limits, of course). For example, the program current can be set with a resistor connecting the programming terminal (collector of Q_1) to ground. This circuit is shown using PNP transistors, but can also work with NPN transistors if the voltages are reversed.



How does this work? $I_\mathrm{C} \approx I_\mathrm{p}$ for Q_1 if we assume a large value of β. Then this sets V_{BE} via the Ebers–Moll equation, and thus V_B for Q_1. The transistor bases are connected, so this also sets V_B for Q_2, and thus V_{BE} for Q_2, since Q_2's emitter is connected to the same supply as Q_1's. Then, since Q_2 satisfies the same Ebers–Moll equation as Q_1, Q_2's collector current must be the same as Q_1's. Note that we didn't need the exact *form* of the Ebers–Moll equation, just that it relates I_C to V_{BE}, and that it is the *same* for both transistors. This is *only* the case if the two transistors are identical and at the same temperature; if these conditions are not true, the analysis is more complicated. In practice, to make sure the properties and temperature match, the current mirror could be implemented using a matched transistor pair in a monolithic package.

Note that we ignored the base current in the above treatment. The base–collector connection on Q_1 shunts the base current for *both* transistors through I_p. Thus, a better expression for the load current is

$$\left(1 + \frac{2}{\beta}\right) I_{\mathrm{load}} = I_\mathrm{p}, \tag{4.74}$$

since the program current is the load current plus two base currents. This correction leads to a 2% or less discrepancy between the program and load currents.

This is a circuit that is used, for example, as a common building block in integrated circuits. For example, the OPA622 op-amp uses an external resistor connected to internal current mirrors to set the **quiescent current** (current when the circuit is idling)—this allows the user to set the trade off between power efficiency and high speed.[2]

4.11.5 Other Refinements to the Transistor Model

We will close out our discussion of BJTs by noting some other complications that are useful to keep in mind.

4.11.5.1 Temperature Dependence of the Base–Emitter Voltage

First, remember that due to the diode-like nature of the BJT, the "input voltage" V_{BE} depends on temperature. We can get idea of the strength of this dependence by getting the temperature slope from the diode law. For example, differentiating the diode law in the form (4.66) gives

$$\frac{dV_{\mathrm{BE}}}{dT} = \frac{V_\mathrm{T}}{T} \log\left(\frac{I_\mathrm{B}}{I_\mathrm{S}} + 1\right) = \frac{V_{\mathrm{BE}}}{T}. \tag{4.75}$$

[2]See `http://www.ti.com/lit/ds/symlink/opa622.pdf`. This is a good exercise: find the current mirrors in the schematic diagram, Fig. 2 p. 10.

Assuming $V_{BE} \approx 0.6\,\text{V}$, this is about $2\,\text{mV}/^\circ\text{C}$ at 25°C. However, this is wrong! It turns out that I_S *increases* exponentially with temperature, which tends to counteract the temperature dependence in V_T. The net effect is that

$$\frac{dV_{BE}}{dT} \approx -2.1\,\text{mV}/^\circ\text{C}, \tag{4.76}$$

or different from the naïve calculation by about a minus sign, and the proportionality to T^{-1} still approximately holds.[3]

4.11.5.2 Early Effect

The **Early effect** says that V_{BE} also depends on V_{CE}. The dependence is weak, and an increase in V_{CE} *decreases* V_{BE} slightly:[4]

$$\Delta V_{BE} \approx -10^{-4}\Delta V_{CE}. \tag{4.77}$$

4.11.5.3 Miller Effect

The **Miller effect** says that the collector–base junction, which normally acts like a reverse-biased diode, acts as if it has a small parallel capacitance, on the order of a few pF. (Remember that a reverse-biased junction has a depletion region, which acts as a thin, insulating layer.) The main problem is that if a transistor circuit has voltage gain G, then this Miller capacitance C_{CB} gets "transferred" to the input as an effective capacitance of $(1 + |G|)C_{CB}$ (Problem 10). With any input impedance, this forms a low-pass filter, so for fast circuits, either the input impedance needs to be kept small, or the gain G must be small.

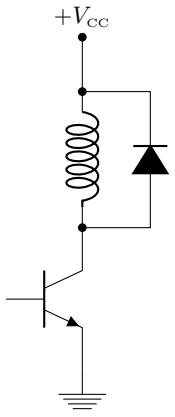
4.11.5.4 Variation of β

Finally, we have already noted that β is not *really* a constant, but this is worth reiterating. It varies between transistors and with temperature. The *only* thing to rely on should be that β is large (100 or more), not that it has any particular value. Note that in terms of gains, β drops out of all the circuits we analyzed; it only appears in the impedance expressions, where its exact value is not critical.

4.12 Circuit Practice

4.12.1 Transistor Switching an Inductive Load

Consider the circuit below, where an NPN transistor switches an inductive load (electromagnet, motor, etc.).



[3] Paul Horowitz and Winfield Hill, *The Art of Electronics*, 2nd ed. (Cambridge, 1989), p. 81 (ISBN: 0521370957).
[4] Horowitz and Hill, *op. cit.*, p. 75.

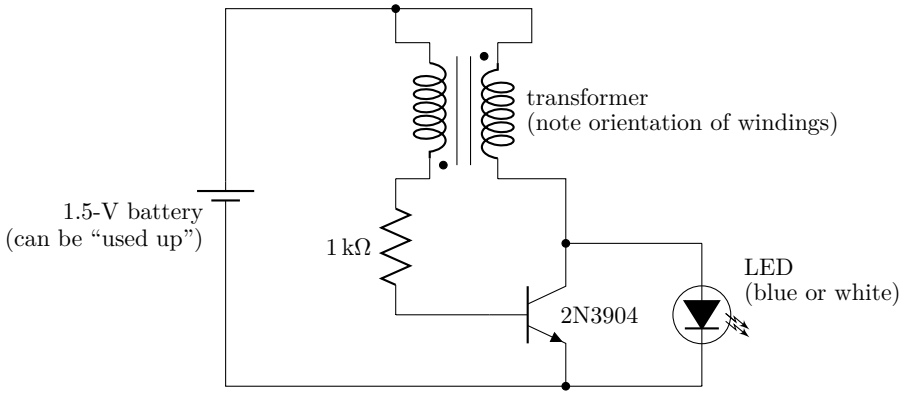
Explain what you need to do to switch the inductive load on and off. Also, why the diode, which is necessary for large inductances, to protect the transistor when switching off the current?

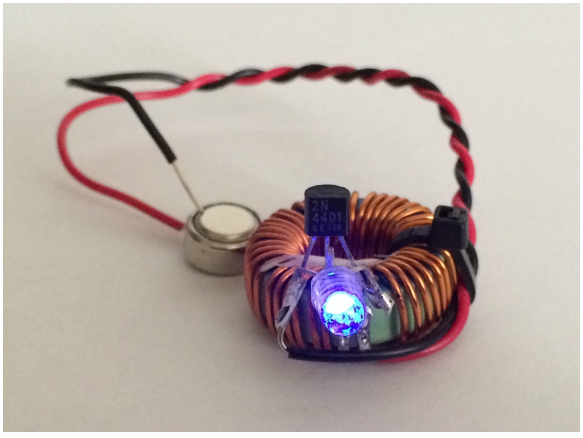
Solution. Without any connection to the transistor base, the inductor is switched off. To run current through the load (say a current I, limited by the resistive part of the load impedance), we need to inject about 1/100 ($1/\beta$) of I into the base, e.g., using a voltage and a resistor as a voltage-to-current convertor.

The problem with this setup is that when switching off the current, the inductor will develop a large EMF to try to sustain the current. This will pull the collector voltage far above $+V_{CC}$ if the inductance is large and the switching is rapid [remember the EMF is $L(dI/dt)$]. This can exceed the transistor breakdown voltage, destroying the transistor. When $V_C > V_{CC}$, the diode shorts the collector to V_{CC}, clamping the voltage at V_{CC} and protecting the transistor.

4.12.2 Joule Thief

Let's take a break from quantitative analysis of transistor circuits, and look at a fun and elegant circuit. The circuit below is called the **joule thief**.[5] To understand the name, consider that the circuit is powered by a 1.5 V battery, but as we discussed before in Section 4.5, turning on a blue or white LED takes about 3.3 V. The "thievery" comes from noting that this circuit works even with a sagging battery with an even lower voltage—hence, we can get steal every last joule from the battery (here, for the purposes of lighting up the LED).





Your exercise here is to trace through the circuit and explain how it works. Just treat the transistor as a switch; no need to do any calculations. Also, as a hint, note that inductors are fairly rare in circuits; usually they show up either as filters (as in the outputs of switching power supplies for computers), or—as is the

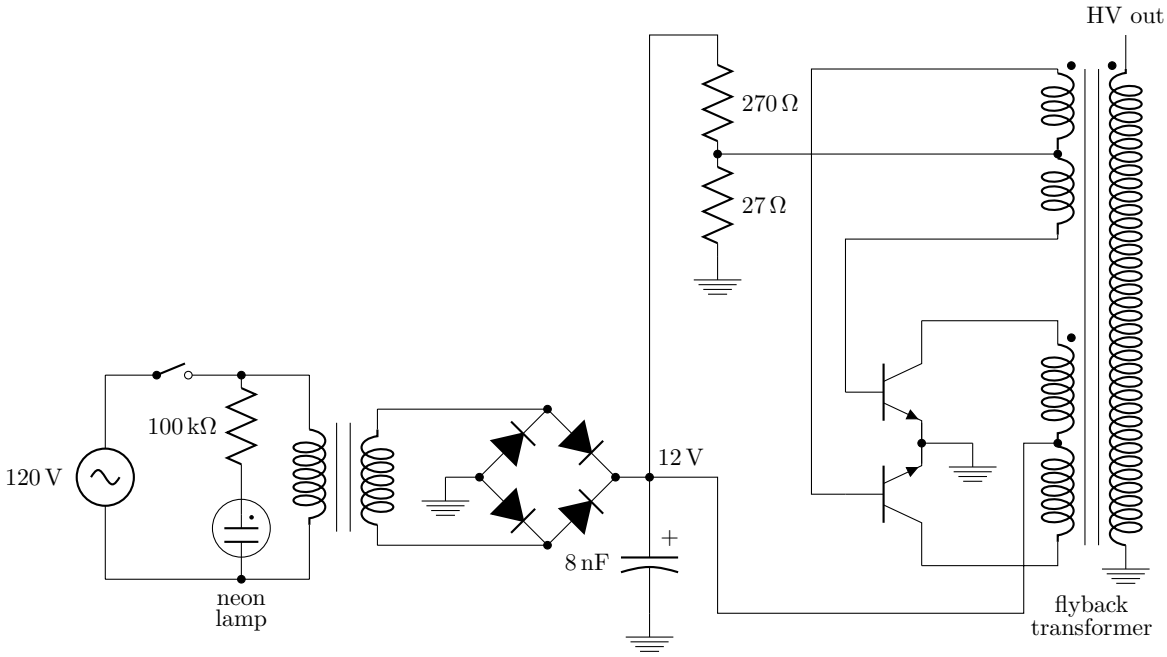
[5] For more on this circuit, including instructions on how to build this, see `http://www.evilmadscientist.com/2007/weekend-projects-with-bre-pettis-make-a-joule-thief/`. See the rest of the site for a lot more cool stuff on electronics and other subjects.

case here—they are around to give an inductive kick when they are interrupted (like the spark-plug coil in gasoline-powered car engines).

Solution. Suppose we start with the transistor off. Current starts flowing from the battery, but it can't go into the collector. The only path it has is into the base; the resistor is there to make sure we don't overdo the current. This base current turns on the transistor, so a much larger current can start to flow into the collector. As the current flows through the secondary of the transformer to the collector, the *opposite* current is induced in the primary (note the primary and secondary are wound/oriented oppositely). This induced current opposes and thus interrupts the base current, switching off the transistor. This interrupts the collector current, interrupting the relatively large current in the transformer secondary. This leads to an inductive kick, and the collector voltage builds up potentially far above the battery voltage (recall the same thing happens for a transistor switching off an inductive load, as in Section 4.12.1). The collector thus builds up to a sufficiently high voltage that the LED turns on, dumping the inductive kick. Then we are back to the initial state, and the process repeats.

4.12.3 Solid-State Tesla Coil

Here is another good circuit to practice *qualitative* reading of transistor operation.



The goal of this circuit[6] is to develop high voltages (of order 25 kV) by using a flyback transformer from a television (the old, cathode-ray-type, which needs large voltages to accelerate the electron beam). The left-hand side of the circuit is basically a dc power supply to drive the right-hand side. The neon lamp (a tiny, neon-filled glass discharge tube) is a handy way of indicating 120-V power; the lamp needs about 90 V to "fire," and then once the discharge starts, little current is needed to sustain it, so the voltage is regulated by the 100-kΩ resistor.

The flyback transformer is driven through a center-tapped primary coil by an alternating pair of power transistors, which are driven by a center-tapped "feedback" coil. Trace through the circuit to understand how the transistor pair switches between the "on–off" and "off-on" pair states. To get started, assume that both transistors are initially off, and pick an arbitrary path for the current from the voltage divider driving the feedback coil.

[6]Robert E Iannini, *Build Your Own Laser, Phaser, Ion Ray Gun and Other Working Space Age Projects* (McGraw–Hill, 1983) (ISBN: 0830606041).

4.12.4 Eric Clapton Signature Stratocaster Preamplifier

Look at the linked schematic diagram[7] for the Eric Clapton Signature Stratocaster from Fender Musical Instruments Corp.[8] The guitar emulates Clapton's beloved old "Blackie" guitar, but the preamp has an unusual feature of enabling a boost in the midrange audio band. This allows electronic control of the guitar's tone from the traditional "Strat" sound to something more akin to a Gibson Les Paul (which featured "humbucking" pickups, which have more midrange gain).

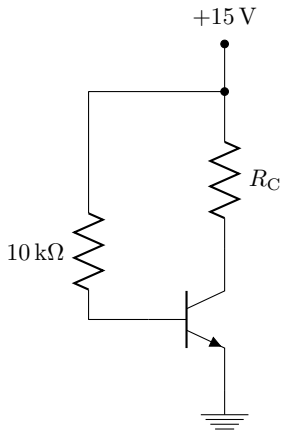
This schematic is somewhat awkward to read, but you can still look through it and try to spot a few elements. In particular, look at transistors Q1–Q4 and identify what amplifiers each one makes up. Also try to identify the bias network in each case.

[7] http://www.blueguitar.org/new/schem/_gtr/ec_schem_fact.jpg
[8] http://www.fender.com/guitars/stratocaster/eric-clapton-stratocaster/product-011760.html

4.13 Exercises

Problem 4.1

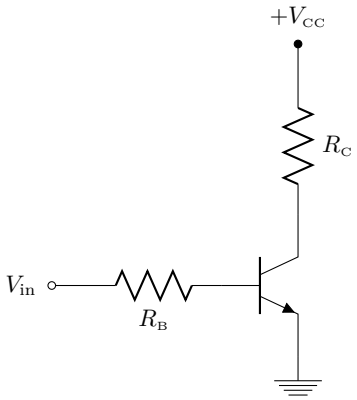
In the circuit below, for what range of resistances R_C will the transistor be saturated?



Assume $\beta = 100$.

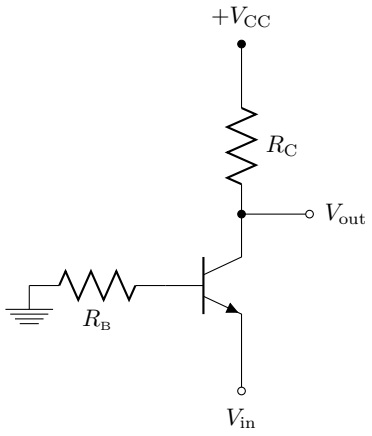
Problem 4.2

Consider the following transistor circuit. Consider V_{CC}, R_B, and R_C to be fixed. For what range of input voltage V_{in} is the transistor saturated?



Problem 4.3

Consider the transistor circuit below. Assume the input is biased such that the transistor works normally, and ignore any internal resistances in the transistor. For the calculations in this problem, you may assume $\beta \gg 1$.



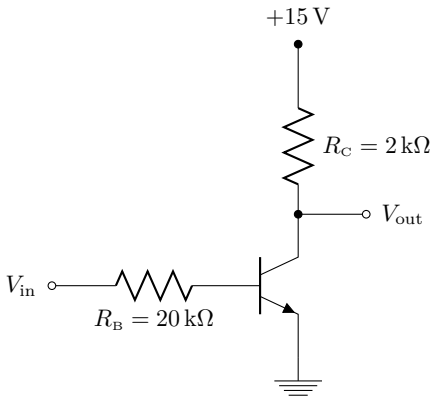
(a) Consider small voltage changes v_{in} and v_{out} at the input and output (i.e., ignore bias voltages). Compute the voltage gain $G = v_{out}/v_{in}$.

(b) Compute the input impedance.

(c) Compute the output impedance.

Problem 4.4

Consider the common-emitter amplifier (in this case, a "grounded-emitter amplifier," since $R_E = 0$) shown below.

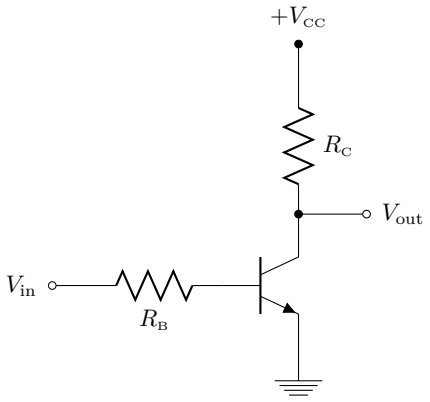


(a) What is the minimum value of V_{in} that saturates the transistor? Assume $\beta = 100$.

(b) What is the (ac) voltage gain of the amplifier, assuming the input is biased to something above 0.6 V and below your result from (a)?

Problem 4.5

Consider the following ampifier.

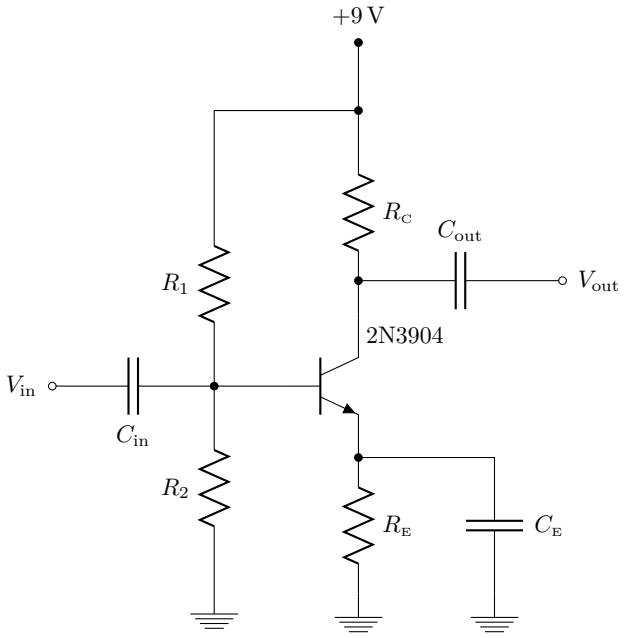


(a) If the input voltage has a dc bias V_0 plus a small ac component $v(t)$ [i.e., $V_{in} = V_0 + v(t)$] over what range of V_0 will the circuit amplify the ac signal? [Write your answer in terms of β, R_B, R_C, V_{CC}, and particular voltages (like 0.6 V).]

(b) Derive an expression for the (ac) gain as a function of R_B, and R_C.

Problem 4.6

Consider the common-emitter amplifier below. This is a fairly involved problem, but the idea is to give you a template for how to design a real transistor amplifier.



The design criteria for this amplifier are: a large gain over a usable bandwidth of 100 Hz–20 kHz, an input impedance of 10 kΩ, a quiescent current (dc current) of 1 mA (dominated by the collector current), and the circuit will drive a 10 kΩ load (the last spec is typical of line-level audio circuits). The dynamic range of the output should also be reasonably close to the maximum possible, so we will fix the collector voltage at $+V_{CC}/2$.

(a) Now we must set the gain of the circuit, and so we need to explain the function of C_E. The idea is that at frequencies above 20 Hz, the capacitor bypasses the resistor, so that over the amplifier bandwidth the intrinsic emitter resistance sets the voltage gain, not R_E. What collector current do we need? Write an expression in terms of R_C, which is yet to be determined.

(b) Use your result from (a) to show that the (ac) voltage gain in this circuit can be written

$$G = -\frac{V_{R_C}}{V_T}, \tag{4.78}$$

where V_{R_C} is the voltage drop across R_C.

(c) Using the specified collector voltage, what is the gain at 20°C?

(d) What are R_C, r_e, and I_B? Use β (h_{FE}) from the data sheet.[9]

(e) For the emitter resistor, a good choice for R_E is often about $0.1 \cdot R_C$. What is this resistance, and what is V_E? The reason for this choice is as follows. Recall that V_{BE} varies with slightly with temperature. If we fix V_B, then this variation leads to a temperature dependence of V_E and thus the bias current. To minimize this effect, V_E should be much larger than the variation in V_{BE}. Verify explicitly that this is the case for this circuit, over, say, variations of 10°C (i.e., estimate the effect on the quiescent current for this temperature change).

(f) What is V_B? (Use the data sheet for any values you need, don't assume standard values.) Now choose R_1 and R_2, accounting for the design specs and not forgetting to account for the effect of R_E on the input impedance. A computer here is probably helpful in solving the resulting equations; ask for help with this if you need it.

(g) Set C_E by guaranteeing that its impedance is negligible compared to r_e (not R_E) over the amplifier's ac bandwidth.

(h) Choose C_{in} and C_{out}.

(i) As a sanity check, note that transistors have a parameter h_{oe}, called the **output admittance**, defined by

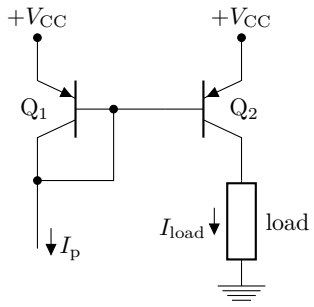
$$h_{oe} := \left.\frac{I_C}{V_C}\right|_{I_B=0}. \tag{4.79}$$

Note that this has units of conductance, and in fact $1/h_{oe}$ acts as an effective collector resistance that appears in parallel with R_C. Verify from the data sheet that we are justified in ignoring this.

(j) Give an order-of-magnitude estimate for the life of a 9 V battery powering this circuit. (Look at some battery data sheets to find some useful information; cite any sources you use.)

Problem 4.7

Consider the current mirror below.

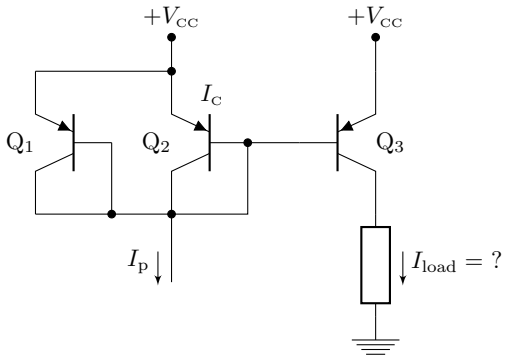


Suppose that Q_1 is at temperature T_1 and Q_2 is at temperature T_2 (in class we assumed these were the same). Derive an expression for the current I_{load} in terms of the program current I_p and the temperatures. Assume that the transistors are otherwise identical.

Problem 4.8

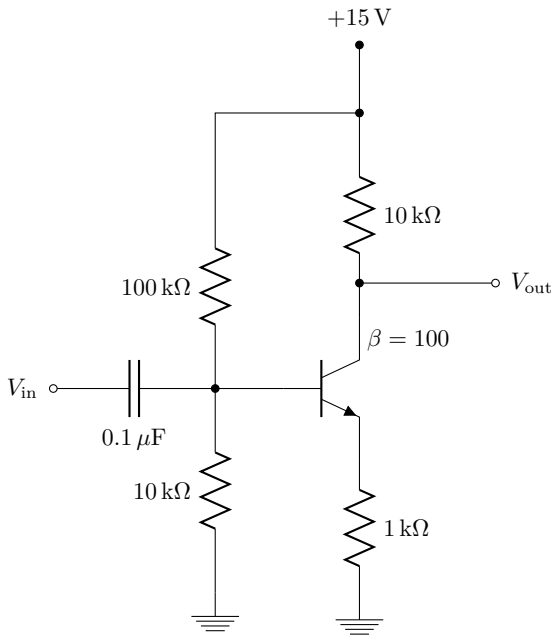
The circuit below is a variation on the transistor current mirror. Compute I_{load}, assuming that the three transistors are identical.

[9] http://www.fairchildsemi.com/ds/2N/2N3904.pdf



Problem 4.9

Consider the circuit below.



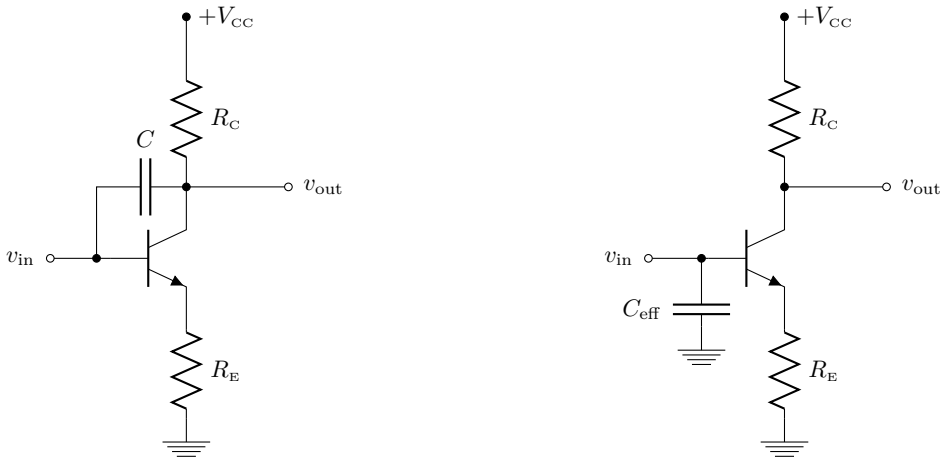
(a) What is the input impedance at high frequencies? Don't bother computing the actual value (unless you really want to), but give enough of an expression to make it clear how you *would* do this if you had a calculator.

(b) What is the dc bias level of V_{out}? Again, you don't need to get the numerical result, just give a chain of relations that shows how you could get the correct numerical answer. Nevertheless, you may want to give a *rough* estimate of the answer as a sanity check and to help in the next part of the problem.

(c) How does the intrinsic emitter resistance enter into the analysis of this circuit? For the specific component values here, how much does it affect the quantities you calculated in (a) and (b)?

Problem 4.10

Consider the common-emitter amplifier shown below, assumed to be biased as needed for the circuit to work normally as an amplifier. Recall that the **Miller effect** refers to an intrinsic capacitance that appears across the reverse-biased base–collector junction, shown as capacitance C in the left-hand diagram.



The reason that the Miller capacitance is a serious problem in a high-gain circuit is that seen as an effective input capacitance C_{eff} as shown in the right-hand diagram, the effective capacitance is

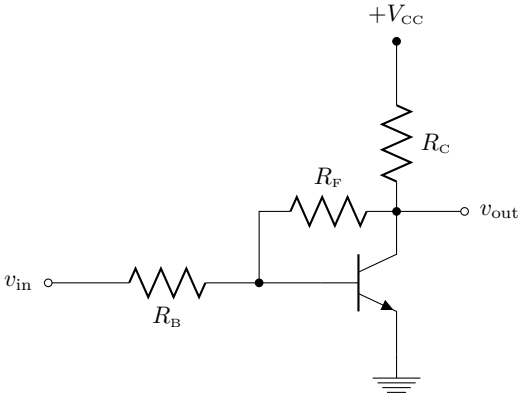
$$C_{\text{eff}} = \big(1 + |G|\big)C, \tag{4.80}$$

where $G \approx -R_{\text{C}}/R_{\text{E}}$ is the ac voltage gain of the amplifier.

Show that this statement is true by considering the change in the charge Q on the capacitor due to a change v_{in} in the input voltage.

Problem 4.11

Compute the ac voltage gain for the transistor amplifier shown below, which is a grounded-emitter amplifier, but with a negative-feedback path via resistor R_{F} from the collector to the base. Assume the input is biased as needed to make the transistor work normally.



Hint: first solve the circuit *without* resistor R_{F} in the circuit, then adapt your solution to the case where R_{F} is present.

Why is the feedback via R_{F} *negative*? (Explain **briefly**.)

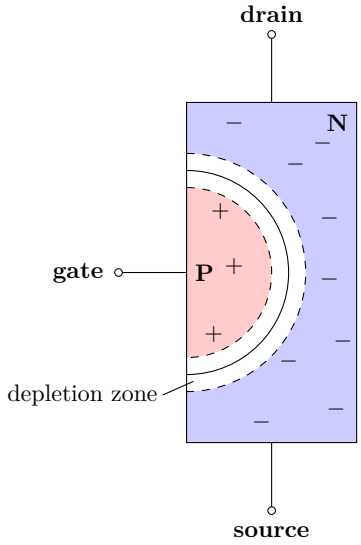
Chapter 5

Field-Effect Transistors and Semiconductor Switching Devices

Field-effect transistors are three-terminal devices, like bipolar junction transistors. Although they operate somewhat differently from BJTs, they can also be used as amplifiers and switches. By considering how they work in basic circuits, we'll see some of the advantages and disadvantages of FETs relative to BJTs. In this chapter we will discuss FETs as well as some related semiconductor devices that are useful as switches, particularly in power-switching applications.

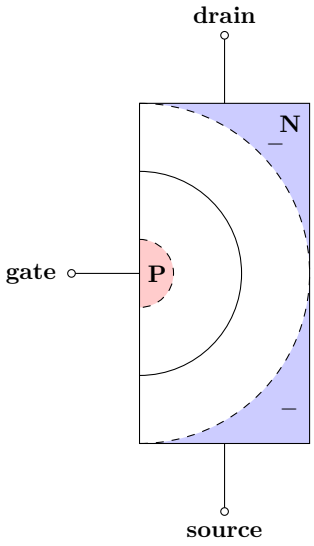
5.1 JFET (Depletion-Mode FET)

A **junction FET (JFET)** is basically a p-n junction with a special geometry and three terminals. JFETs always work as **depletion-mode** devices, which refers to the depletion zone at the junction, which is responsible for the switching action of the JFET. The basic scheme for an **n-channel JFET** is shown below (the **p-channel** counterpart is basically the same, under the exchange of p- and n-type semiconductors).



As in the semiconductor diode, a depletion zone forms at the p-n junction. In normal operation, the voltage at the gate terminal is kept below that of the drain and source, so the junction is reverse-biased. Only a small leakage current flows via the gate terminal. However, in the configuration shown, there is a low-impedance path between the drain and source terminals, through the n-type region (hence, the "n-channel").

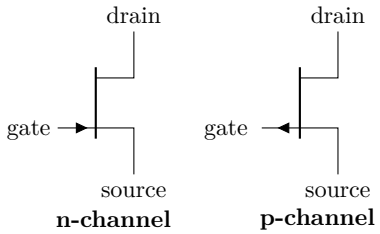
However, when the gate is brought to a negative voltage with respect to drain and source, the depletion zone expands as in the reverse-biased diode. At a sufficiently negative voltage, the depletion zone "pinches off" the n-channel, preventing current flow from drain to source.



For intermediate gate voltages, the gate voltage acts as a control that modulates the resistance of the drain-source path. This acts something like modulating the flow of water in a garden hose by changing the clamping force of a pair of pliers on the hose.

Note that the convention in the n-channel JFET is that **current flows from drain to source** (i.e., the n-type carriers are flowing from the source and out the drain). The current flows from source to drain in the p-channel JFET. These two terminals appear to be interchangeable according to the above diagrams, and for some devices they are so in practice. However, due to geometric differences, the drain and source are not always equivalent. For example, the drain and source often have different capacitances, so reversing them can affect the speed of a JFET amplifier.

The symbols for n-channel and p-channel JFETS is shown below. The difference between the two transistors is only the direction of the gate arrow, which indicates the orientation of the p-n junction (like the arrow in the diode symbol).



The asymmetric placement of the gate differentiates the drain and source.

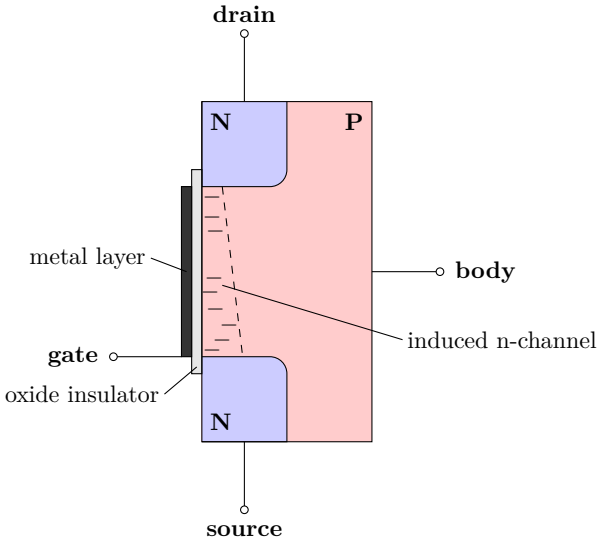
To summarize the operation of the n-channel (depletion-mode) JFET:

- If $V_{GS} = 0$, then current can flow through the n channel, typically from drain to source (i.e., $I_{DS} > 0$).
- If $V_{GS} < 0$, the junction is reverse-biased, and the expanding depletion zone restricts current flow.
- There is some **threshold voltage** V_T: if $V_{GS} < V_T$, then the n-channel is "pinched off," and no current flows ($I_{DS} = 0$). Typically V_T ranges from -2 to $-15\,\text{V}$. *Don't* confuse the threshold voltage from the thermal voltage from diode-law and Ebers–Moll fame.
- The forward-biased case $V_{GS} > 0$ doesn't normally happen. You may as well just use a diode.

- The JFET is thus a **transconductance device**, where a voltage controls a current (like the BJT, from the standpoint of the Ebers–Moll equation).

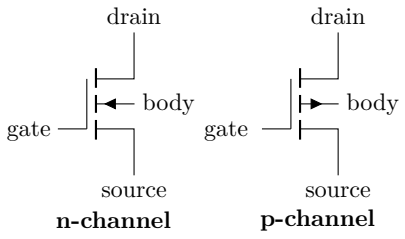
5.2 MOSFET (Enhancement-Mode FET)

A second major class of FETs is the **metal-oxide-semiconductor FET (MOSFET)**, also called the **insulated-gate FET (IGFET)**. The names refer to the gate terminal, which is connected to a metal conducting layer, and insulated by an oxide layer from the semiconductor regions of the MOSFET. Some MOSFETs behave as depletion-mode devices, but these are relatively rare. Here we will focus on the much more common case of **enhancement-mode MOSFETs**. The basic scheme for an n-channel MOSFET is shown below (note that the body is a p-type semiconductor, unlike the n-channel JFET; the operating principle is quite different).



The drain and source terminals connect to n-type regions, which are embedded in the p-type substrate or body. In principle MOSFETs have a separate body connection, but these are not always explicitly available (often the body and source terminals are combined into one terminal). The idea here is that if no voltage is applied to the gate, no drain–source current can flow because it will be blocked by one reverse-biased p-n junction. However, when a *positive* control voltage (with respect to drain and source) the gate's E-field pulls *n-type* carriers out of the p-type substrate (the p-type carriers are the **majority carrier** in the p-type semiconductor, but n-type carriers are also present as the **minority carrier**). The n-type carriers bunched against the gate insulator form an effective n-channel, or **induced n-channel**, that bridges the drain and source. The diagram shows the induced n-channel as being somewhat asymmetric, as appropriate if $V_{\text{DS}} > 0$.

The symbols for n-channel and p-channel enhancement-mode MOSFETS are shown below. Again, there are extra body connections, but often these are internally shorted to the source.



Also, the three short, vertical lines that represent the three semiconductor regions are sometimes drawn as a single line. Again, the asymmetry of the gate distinguishes drain from source, and the orientation of

the arrow distinguishes n- and p-channel types, by indicating the orientation of the p-n body–drain and body–source junctions.

To summarize the operation of the n-channel (enhancement-mode) MOSFET:

- If $V_{GS} = 0$, no current I_{DS} can flow, because of a reverse-biased junction as in the NPN transistor.
- If $V_{GS} > 0$, the gate field induces an n-channel, and current I_{DS} can flow, typically with $I_{DS} > 0$.
- Due to the oxide insulator, there is very little gate-current leakage. However, the insulator layer is easily damaged by static discharges.
- The switching operation is similar to the n-channel JFET, but the threshold voltage $V_T > 0$. That is, conduction occurs when $V_{GS} > V_T$, so the gate voltage must be *positive* for the MOSFET. Remember the gate voltage is generally *negative* in the JFET, with conduction turning off for sufficiently negative voltages.

MOSFETs are very common in digital circuits, in the form of **complementary MOS (CMOS)** circuits, where n- and p-channel MOSFETs are paired together.

5.3 Quantitative FET Behavior

Having established that JFETs and MOSFETs have similar behavior except for the particular value (i.e., sign) of the threshold V_T, we can treat both cases together, so long as we track the control voltage V_{GS} *relative to* V_T. For small input signals (small ac signals on a dc bias), the transconductance nature of a FET means we can write

$$i_{DS} = g_m v_{GS}, \tag{5.1}$$
(FET transconductance relation)

where g_m is the **transconductance**. This has dimensions of Ω^{-1}, which is often written $\mho$ and called a **mho** (or **siemens**, abbreviated "S," if you want to be all SI about it). The transconductance depends, however, on the bias levels V_{GS} and V_{DS}. This relation is the analogue of $I_C = \beta I_B$ for BJTs.

The FET current formulas can describe the current-voltage characteristics, like the Ebers–Moll equation for BJTs, but in a piecewise way.

1. **Linear region:** when $V_{DS} < (V_{GS} - V_T)$, we have

$$I_{DS} = 2k\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{5.2}$$
(FET in linear region)

where k is a conductance parameter, which is device-dependent and scales with temperature as $T^{-3/2}$. (The threshold voltage V_T also depends on temperature.) Note that due to the quadratic term in V_{DS}, this relation is not really "linear." However, this term is negligible if $V_{DS} \ll (V_{GS} - V_T)$, or if more linear behavior is desirable, there are tricks to compensate for this term. In the really linear case where we can ignore the nonlinear term, we have a resistance

$$R = \frac{1}{2k(V_{GS} - V_T)}. \tag{5.3}$$
(FET resistancein linear region)

This says that the FET in this regime is useful as a voltage-controlled resistor, for example to control variable gain or attenuation in a circuit.

Saturation region: when $V_{DS} > (V_{GS} - V_T)$. In this region, I_{DS} is independent of V_{DS}, in some sense like the saturated BJT where the transistor no longer directly modulates the collector current. In this region,

$$I_{DS} = k(V_{GS} - V_T)^2. \tag{5.4}$$
(FET in saturation region)

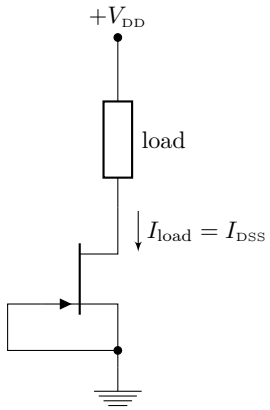
Subthreshold region: when I_{DS} is small, because the control voltage is around or below the threshold. In this region,

$$I_{DS} = k\, e^{V_{GS} - V_T}. \tag{5.5}$$
(FET in subthreshold region)

5.4 Basic FET Circuits

5.4.1 JFET Current Source

A simple JFET circuit is the JFET current source, show below.



The idea is to operate the JFET in the saturation region, where I_{DS} is independent of V_{DS}, allowing the JFET to adjust the load voltage to regulate its current. Since the gate and source are shorted (and grounded), $V_{GS} = 0$. Then using Eq. (5.4), we find the constant current

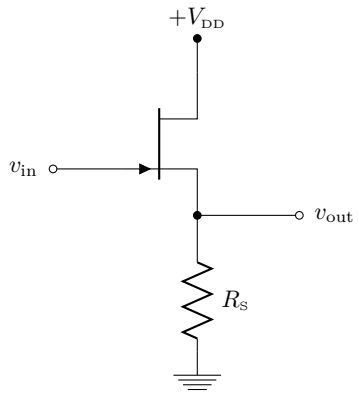
$$I_{DS} = kV_T^2 =: I_{DSS}, \tag{5.6}$$
(JFET regulated current)

where I_{DSS} is the maximum JFET current.

The advantage of this circuit, compared to the BJT current source, is its simplicity. The disadvantage is more serious: I_{DSS} varies significantly between devices. However, it is possible to get hand-picked FETs with particular values of I_{DSS}, called **current-regulator diodes** (something like Zener diodes, but for current instead of voltage).

5.4.2 JFET Source Follower

The next circuit is the JFET analogue of the BJT emitter follower.



Ignoring offsets, Ohm's law for the source resistance gives

$$v_{\mathrm{S}} = i_{\mathrm{DS}} R_{\mathrm{S}}, \tag{5.7}$$

while the transconductance relation (5.1) gives

$$i_{\mathrm{DS}} = g_{\mathrm{m}} v_{\mathrm{GS}} = g_{\mathrm{m}}(v_{\mathrm{G}} - v_{\mathrm{S}}). \tag{5.8}$$

The latter equation becomes

$$v_{\mathrm{G}} = \frac{i_{\mathrm{DS}}}{g_{\mathrm{m}}} + v_{\mathrm{S}} = \frac{i_{\mathrm{DS}}}{g_{\mathrm{m}}} + i_{\mathrm{DS}} R_{\mathrm{S}}. \tag{5.9}$$

after solving for v_{G} and using Eq. (5.7). Then the ac voltage gain is

$$G := \frac{v_{\mathrm{out}}}{v_{\mathrm{in}}} = \frac{v_{\mathrm{S}}}{v_{\mathrm{G}}} = \frac{i_{\mathrm{DS}} R_{\mathrm{S}}}{i_{\mathrm{DS}}/g_{\mathrm{m}} + i_{\mathrm{DS}} R_{\mathrm{S}}}, \tag{5.10}$$

after using Eqs. (5.7) and (5.9). This simplifies to

$$G = \frac{g_{\mathrm{m}} R_{\mathrm{S}}}{1 + g_{\mathrm{m}} R_{\mathrm{S}}}. \tag{5.11}$$

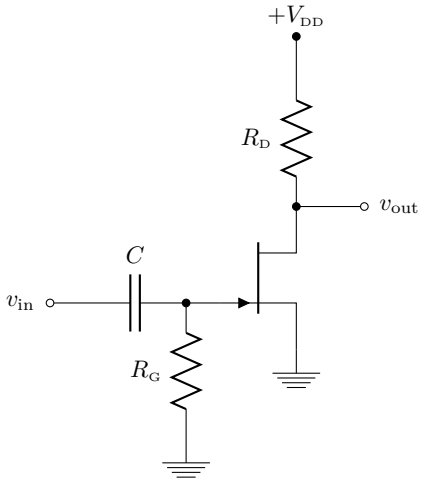
(voltage gain, JFET source follower)

Note that $G \approx 1$ (hence, a follower) for large resistance $R_{\mathrm{S}} \gg 1/g_{\mathrm{m}}$. For a typical signal JFET like the 2N5485, $g_{\mathrm{m}} \approx 5000\,\mu\mho$, so $1/g_{\mathrm{m}} \approx 200\,\Omega$, so have a larger resistance than the internal FET resistance is not difficult.

This circuit has high input impedance, because the input is essentially a reverse-biased diode. However, the output impedance is basically $1/g_{\mathrm{m}}$, which could be somewhat large compared to the emitter follower. Another disadvantage of this circuit is the unpredictable dc offset, since V_{GS} for a certain current is not well-controlled in the fabrication process.

5.4.3 JFET Voltage Amplifier

Next is a JFET voltage amplifier, the FET analogue of the common-emitter amplifier.



Here, R_{G} maintains the dc bias of V_{GS} at ground, but allows ac modulation of the gate via the capacitor C.

If $V_{\mathrm{GS}} = 0$, the quiescent current is $I_{\mathrm{DS}} = I_{\mathrm{DSS}}$, as in the JFET current source (thus assuming that V_{DD} is large enough to put the JFET into saturation). Then the output is biased at

$$V_{\mathrm{out}} = V_{\mathrm{DD}} - I_{\mathrm{DS}} R_{\mathrm{D}}. \tag{5.12}$$

This bias level may be adjusted by introducing a source resistor, thus lowering V_{GS} and decreasing I_{DS}.

For small ac signals,

$$v_{\text{out}} = v_{\text{D}} = -i_{\text{DS}} R_{\text{D}}. \tag{5.13}$$

Then using the transconductance relation (5.1),

$$i_{\text{DS}} = g_{\text{m}} v_{\text{GS}} = g_{\text{m}} v_{\text{in}}, \tag{5.14}$$

we have

$$v_{\text{out}} = -g_{\text{m}} R_{\text{D}} v_{\text{in}}, \tag{5.15}$$

which implies a voltage gain

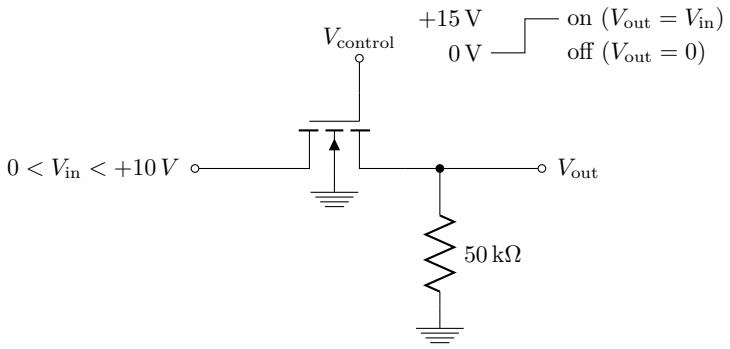
$$G = -g_{\text{m}} R_{\text{D}}. \tag{5.16}$$

(voltage gain, JFET voltage amplifier)

So for example, if $R_{\text{D}} = 1\,\text{k}\Omega$ and for the 2N5485, $1/g_{\text{m}} = 200\,\Omega$, then $G = -5$, which is not a huge gain. By contrast, in the common-emitter amplifier, using the intrinsic emitter resistance $r_{\text{e}} = 25\,\Omega$ at $I_{\text{C}} = 1\,\text{mA}$, the gain is about 8× larger (because r_{e} is eight times smaller than $1/g_{\text{m}}$). So the advantage of the JFET is high input impedance (good for input stages of amplifiers, especially op-amps), the advantage of the BJT is better speed (due to lower capacitance) and amplification.

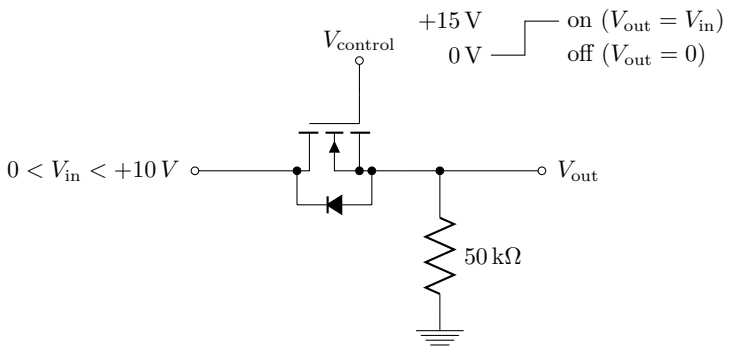
5.4.4 MOSFET Analog Switch

One nice example of a MOSFET application is as an **analog switch**, which passes or blocks an analog signal based on a control voltage.



Here, the base terminal is tied to ground; the MOSFET acts as a short if V_{control} is well above any input, but acts as an open circuit if the gate drops to zero. The output resistor ensures a zero (not floating) output when the MOSFET is off, and it suppresses any tendency of the signal to cross the transistor in the off state due to drain–source capacitance. This circuit can be adapted to positive/negative signals by dropping the base and the "off" voltage to −15 V.

MOSFETs with a separate base connection (like the obsolete 2N4351) are actually the exception, rather than the rule. More common are MOSFETs with the base and source shorted together. The net effect of this connection is similar to connecting a diode from source to drain. The circuit still works with this kind of MOSFET (e.g., 2N7000), if the orientation is correct.



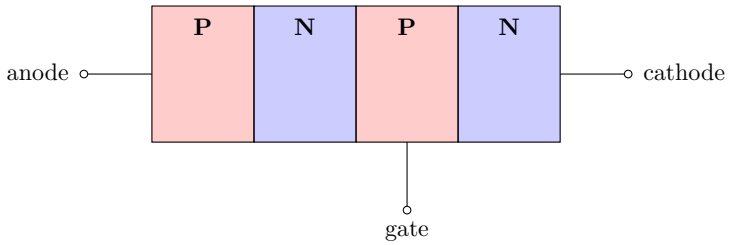
However, for a positive/negative signal, *two* 2N7000's, back-to-back, are needed for the switch to function properly.

5.5 Thyristors

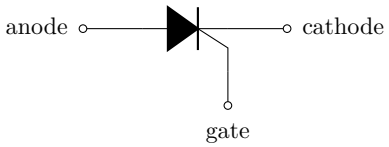
Now we will move on to a few semiconductor components that are somewhat more complicated than the basic transistors (BJT and FET). **Thyristors** refer to a family of devices based on a three semiconductor junctions, though the name "thyristor" can also be used synonymously with the basic SCR component that we will treat first. These are devices that utilize p-n junctions to produce hysteresis, which is useful in switching circuits.

5.5.1 SCRs

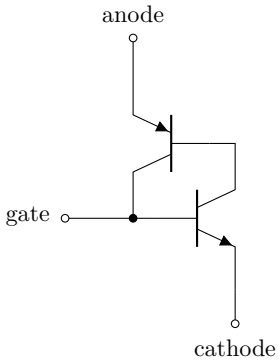
The basic form of a thyristor is the **semiconductor-controlled rectifier**, or **SCR**. An SCR is constructed from alternating four regions. The result is something like a diode, but with two p-n junctions and one n-p junction, all in series.



The basic operation will be something like a diode, but there is one additional "gate" connection to the internal p-type region. This is reflected in the schematic symbol, shown below.



Thinking of the p-n junctions as diode-type junctions is only useful in understanding the "`OFF`" state; in either direction there is a reverse-biased diode, so no current can flow. However, to understand the switching behavior of the SCR, it helps to think of the p-n-p stack to the left as a PNP transistor, and the n-p-n stack to the right as an NPN transistor. Since the inner two regions are common to each transistor, the transistors are interconnected as shown in the equivalent circuit below.



Again, if no current is initially flowing in the circuit, then both transistors are `OFF`, and no current can flow from anode to cathode, because the current would have to pass via the collector of one or the other transistor.

However, if a (sufficiently large) anode–cathode current is already flowing in the circuit, then the collector current of the NPN transistor ensures the PNP transistor is ON (since the current comes via the emitter–base junction), and the collector current of the PNP transistor flows through the base–emitter junction of the NPN transistor, ensuring that it is ON.

As an aside, note that although this transistor model is a useful way to understand how an SCR works, an SCR can't necessarily be replaced by two transistors. For example, the collector leakage current of one of the transistors could supply sufficient base current to turn the other ON, in which case this SCR OFF state wouldn't work as described above.

So now we see that there are two possible states, where current is flowing or not, and the states are self-sustaining (the current-flowing state of course requiring a sufficient forward voltage). Then the question is, how can we transition between the states? In particular, if no current is flowing, how can we get the current started? There are two useful answers to this in practice:

1. Inject some gate current (flowing from gate to cathode). This turns on the NPN transistor, which turns on the PNP transistor, and the SCR transitions to the conducting state.

2. Bring the forward voltage sufficiently high that the reverse-biased junction breaks down. Just before the breakdown, the forward voltage can be quite high (hundreds of volts), but after breakdown, the forward voltage will become much smaller, on the order of a volt. Note that a gate current will reduce the breakdown voltage, and a gate current above a switching threshold brings the breakdown voltage to essentially zero.

To make a transition in the other direction, or that is, to stop the conduction, the only way is to bring the forward current to zero (or equivalently, drop the forward voltage to zero). This "resets" the SCR into the non-conducting state, assuming there is no gate current is off.

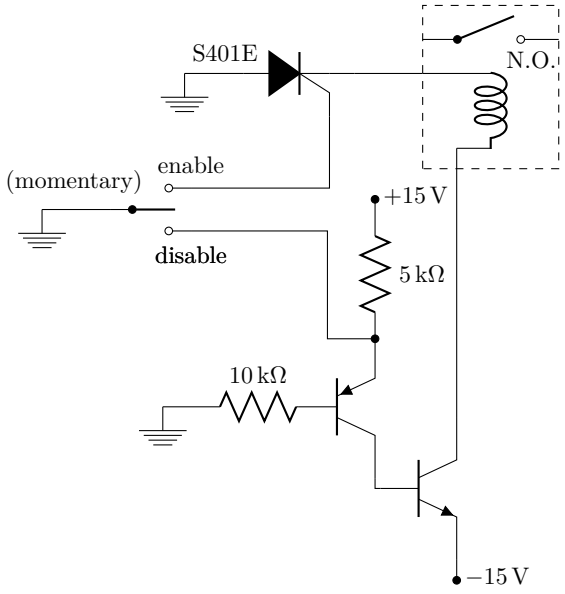
As an example, the small-signal S401E SCR has 1-A maximum (rms) current, 400-V off voltage (in either direction), a gate trigger current guaranteed to be in the range of 1–10 mA, a 1.6-V maximum forward-conduction voltage, and a maximum gate forward voltage of 1.5 V.

SCRs are useful in switching circuits where some "memory" (latching behavior) is required, and they are also useful in switching large currents and voltages. The key to the utility in switching high-power loads is that the forward voltage drop is small in the conducting state, so large currents do not necessarily result in large power dissipation. It is also possible to break down the SCR in the reverse direction (just like a diode) with a sufficiently high voltage, but then the voltage gets clamped to the (high) breakdown voltage, as in a zener diode. In that case, a large current could result in a large power dissipation, because both V and I are large.

The name "thyristor" derives from the tube analogue of the SCR, the **thyratron**, which is a gas-filled tube where a control voltage can initiate an ionization breakdown of the gas, putting the tube into a conducting state (like the neon lamp, discussed below with DIACs). The thyristor name is then a combination of thyratron + transistor.

5.5.1.1 Example: Latching Switch with Power-Supply Fault Protection

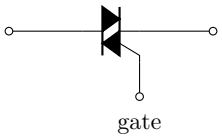
One example of a useful switching circuit is shown below. The idea here is that sometimes one has to switch a delicate and/or expensive load, and the driving circuitry must include some protection against certain faults, such as a failed power supply. (Power supplies can fail due to faulty capacitors, and capacitors elsewhere in the circuit, such as the bypass capacitors discussed in Section 7.6.2.2, can fail to a short and effectively cause a power-supply failure.) The circuit below is intended to switch a delicate load via a relay (magnet-controlled switch). The diagram shows the relay, but not the load, and the relay is normally open (N.O.) when there is no magnet current.



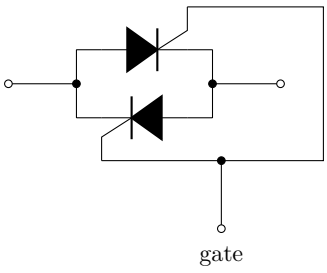
A momentary switch allows the user to toggle the relay on or off. In the "enable" position, the switch activates the SCR, but only if the ±15 V power supplies are working, so that there is a path to −15 V via the transistor pair. Once the SCR is activated, current flows through the relay coil, connecting the load. The "disable" switch position forces the PNP emitter to ground, which turns off the PNP and thus the NPN transistor opening the relay and resetting the SCR. (The disable operation also emulates the failure of the +15-V power supply.)

5.5.2 DIACs and TRIACs

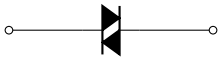
There are two related devices in the thyristor family: **DIACs** and **TRIACs**. Since SCRs are diode-like in their asymmetric operation, it is useful to have analogous devices to work in ac circuits. The TRIAC is basically an SCR that can conduct (with low forward voltage) in either direction. The TRIAC symbol is shown below.



This component is basically equivalent to two SCRs, oppositely oriented and wired in parallel, with a common gate, as shown below.



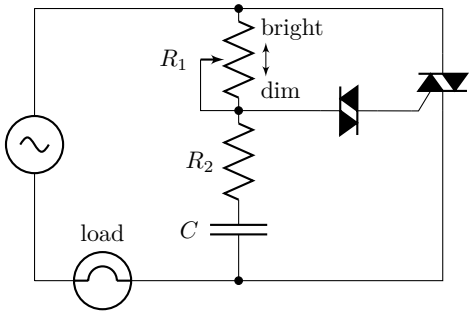
The idea is similar with a DIAC, but this component has no gate; the intent is for this device to conduct by bringing the voltage above threshold (in either direction), to break down the device into conduction.



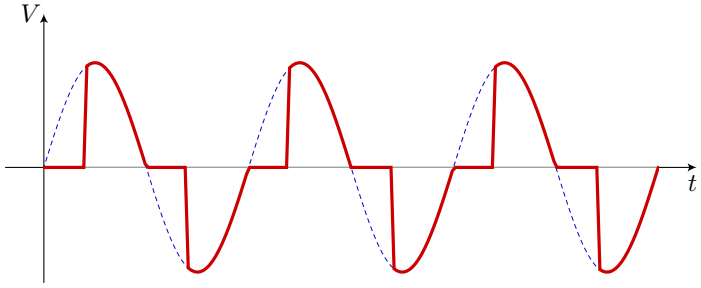
While it's useful to think of both devices as a combination of two SCRs, they are fabricated in a simpler way as a single stack of junctions. For example, the DIAC, because it doesn't need a gate connection, can be fabricated as a p-n-p stack. The symmetry of the stack allows for symmetric-breakdown operation. DIACs are typically fabricated with breakdown voltages in the range of 32–40 V. Gas-discharge lamps can also work in a similar way to DIACs. For example, the small NE-2 neon lamp has a breakdown voltage of 90 V (i.e., it does not conduct until the voltage exceeds 90 V in either direction), and the discharge-state (i.e., lit) voltage drops to around 60 V when the lamp conducts. The nominal conduction current is about 0.5 mA, so if used as a lamp across 120 V ac, a dropping resistor of around 100 kΩ is necessary to limit the lamp current. However, for switching applications, a neon lamp can function in place of a DIAC, for example in the dimmer circuit to follow.

5.5.2.1 Light Dimmer

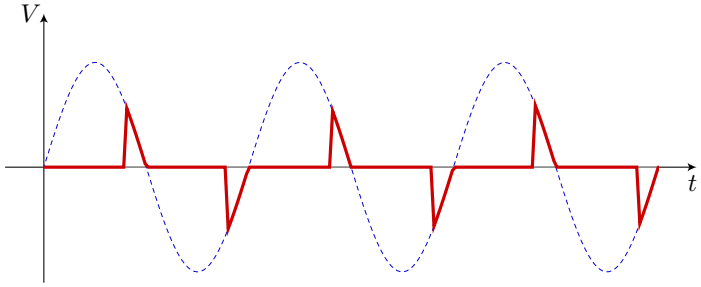
One of the most important applications of thyristor devices is in dimmers for ac lighting circuits. There are a number of designs, but a common DIAC–TRIAC design is shown below.



As each "pulse" (of either sign) from the ac source begins, the TRIAC is initially off, and the voltage begins to charge the capacitor via $R_1 + R_2$. Once the capacitor charges to the breakdown voltage of the DIAC (around 30 V), it conducts, and discharges the capacitor via R_2 through the TRIAC gate, causing it to conduct. The TRIAC continues to conduct until the end of the pulse, supplying current to the load during the last part of the cycle, and the capacitor also discharges and resets before the next voltage pulse. The net result is that the load is only on for part of each cycle; the voltage presented to the load is illustrated below.



The brightness can be adjusted by changing R_1: a larger R_1 leads to a longer charging time, which leads to a longer delay until the load turns on each cycle. A resulting dimmer signal is illustrated in the plot below.



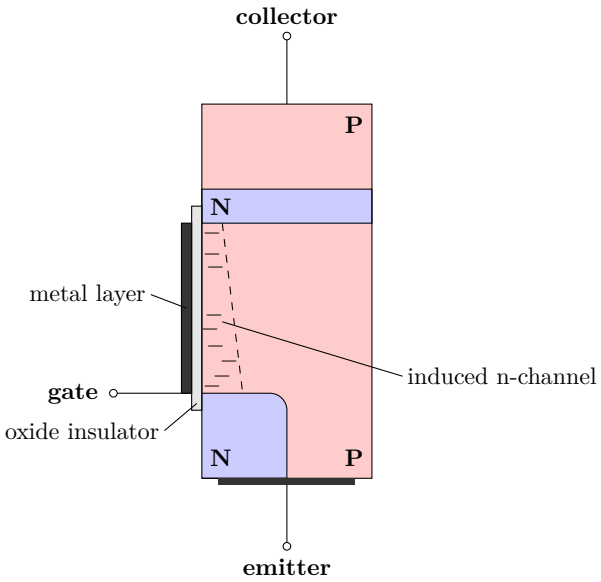
The R_2 resistor protects against excessive capacitor charging or discharging currents.

One disadvantage of this circuit is the sudden turn-on edge in each cycle (in the voltage as well as the current), which is largest at 50% brightness. This can be a source of electromagnetic interference, and it can also cause buzzing in the load (for example, causing mechanical motion of the filaments in incandescent bulbs). Good dimmer designs include inductors in series with the load to suppress the interference and smooth the hard edges.

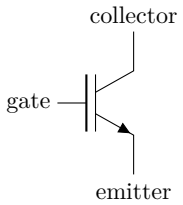
This circuit must also be designed with a particular load in mind (or range of loads), because the load must draw sufficient current to keep the TRIAC on until it resets at the end of the cycle. Thus, for example, older thyristor dimmers designed for incandescent loads can perform poorly when the lights are replaced by lower-current LED equivalents, which draw much less current. The resulting problems typically come in the form of flicker in the new lighting. The other problem that can crop up in this situation is that during the charging cycle, a small current flows through the load from the voltage source. Generally R_2 limits this to something small enough that an incandescent light is effectively off. But with higher-efficiency LED lights in an older dimming fixture, the LEDs can stay dimly lit, even in the dimmest setting (or even the "`OFF`" state of the fixture).

5.6 IGBTs: Switching Very Large Voltages and Currents

The **insulated-gate bipolar transistor (IGBT)** is a device fabricated as shown below. It is something like the n-channel (enhancement-mode) MOSFET (see the diagram on p. 103), but with an extra p-type region, and different names for the conduction terminals.

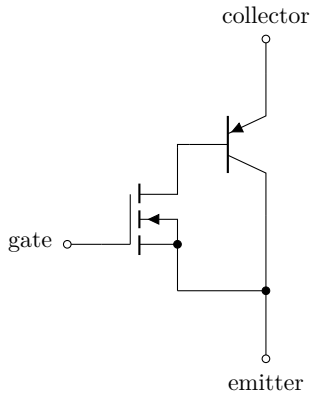


The schematic symbol is shown below: it is basically the symbol for a BJT, but with a gate separation to emphasize the insulated gate. Note that while MOSFETs and BJTs come in n- and p-flavors, the IGBT comes in only the polarity shown (or, that is, the opposite polarity is so rare as to be effectively nonexistent).



Ignoring the upper p-type region, this is just a MOSFET with source and base shorted together. The extra

p-type region then forms something like an PNP BJT with the upper n-type and lower p-type regions. The MOSFET drain and BJT base are connected (because they are formed by the same n-type region), and the MOSFET body/source and BJT emitter are connected. Thus, an equivalent circuit is the transistor pair as connected below.



This circuit functions more or less as a BJT, except the C–E conduction is controlled by a voltage (V_{GE}), rather than a base–emitter current. When V_{GE} is sufficiently positive, the MOSFET turns `ON`, which switches on the PNP transistor and allows current to flow from collector to emitter. Note that in principle this current may flow via either the PNP collector (which is *not* the IGBT collector!) or via the MOSFET drain–source path, but the device is fabricated such that most current flows through the collector. For large currents, this arrangement can be efficient compared to a power BJT, which may require substantial base current to control the transistor (around 10% of the collector current for good saturation, since power BJTs are on the low end of the β spectrum).

For high-power switching applications, power MOSFETs are also available. However, for switching loads with hundreds of amps and hundreds of volts, IGBTs tend to be the best choice in terms of power dissipation. Recall that if the gate voltage is sufficiently large, the MOSFET behaves like a variable resistor. For example, the Infineon IRF1405 power MOSFET is specified to have a typical `ON` resistance of 4.6 mΩ, a maximum continuous drain current of 169 A (at 25°C), and a drain–source breakdown voltage of at least 55 V.[1] In this case, when the MOSFET switches the power to a load, the power dissipated by the MOSFET is $P = I^2 R_{DS}$ while the transistor is `ON`. Since the conduction path for an IGBT behaves like a BJT, the collector–emitter voltage is relatively small when the transistor is saturated, and the IGBT `ON` power dissipated is $P = IV_{CE}$. A comparably sized example of an IGBT (in the sense of being available in a package of similar size) is the Fairchild FGA60N65SMD,[2] which is rated to switch 120 A (at 25°C) at 650 V, with $V_{CE} = 1.9$ V typical.

Since the MOSFET power scales quadratically with current while the IGBT scales only linearly, the IGBT can have a big advantage for high-power loads. For the two example devices above, we can estimate a power dissipation of 228 W for the IGBT vs. 66 W for the MOSFET at 120 A. For these "small" currents, the MOSFET is the clear winner (though the IGBT can of course switch much higher voltages). However, the crossover to where the IGBT has the advantage occurs at $I^2 R_{DS} = IV_{CE}$, or $I = V_{CE}/R_{DS}$, which comes out to about 400 A, if we extrapolate assuming the same device parameters carry over to larger devices. This is a typical figure, above which IGBTs are preferred, especially to stand off large voltages. Of course, if MOSFETs are preferable for some reason, they can be used in a parallel transistor bank to reduce the effective resistance. However, in MOSFETs capable of standing off larger voltages, the conduction path must be longer, typically leading to larger `ON` R_{DS}, and thus worse power dissipation at a particular current.

Such large IGBTs certainly exist, in "mini-brick" and "brick" packages, with screw terminals. For example, the somewhat larger IXYN80N90C3H1 IGBT by IXYS[3] comes in a SOT-227B "miniBLOC" package, with 4 screw terminals . (Two terminals are emitter terminals, so one can function as a "Kelvin emitter."

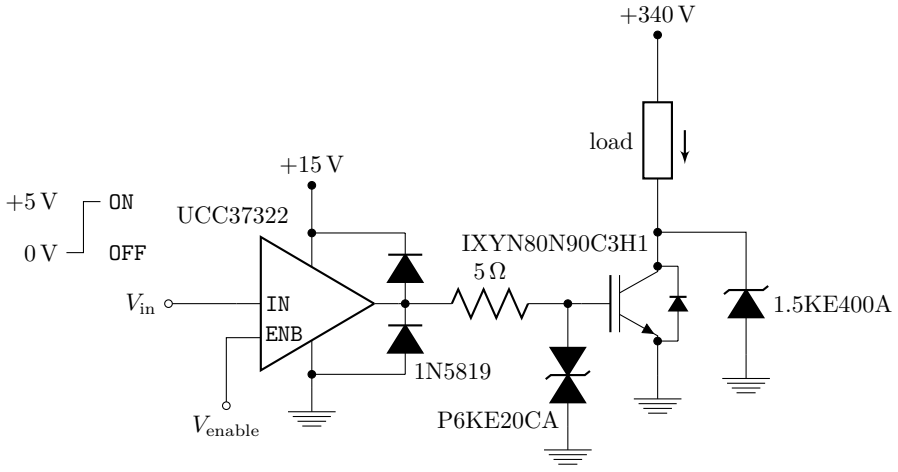
[1] `http://www.infineon.com/dgdl/irf1405pbf.pdf?fileId=5546d462533600a4015355db084a18bb`
[2] https://www.fairchildsemi.com/datasheets/FG/FGA60N65SMD.pdf
[3] http://ixapps.ixys.com/DataSheet/DS100522A(IXYN80N90C3H1).pdf

The second terminal can connect to a voltage-sensing circuit, such as a gate driver, to accurately sample the emitter voltage without a spurious drop across a high-current contact resistance at the other emitter connection to the current load.) This is designed to switch a pulsed current of 340 A (or 115 A continuous) at 900 V, with $V_{CE} = 2.3$ V. The more expensive ($\sim$ \$200) CM600HA-24A by Powerex[4] is rated for 600 A (dc, 1200 A pulsed) at 1200 V, with $V_{CE} = 3$ V.

IGBTs have a reputation for being somewhat slower than MOSFETs for switching, because the main conduction path is governed by the carriers injected into the "base" n-type region. However, the carriers enter and leave this region only indirectly, via the neighboring p-type regions. This is in contrast to a standard BJT, where carriers can enter or leave directly via the base connection. Especially when switching off, excess carriers can be "stuck" in the base region, only escaping by causing collector–emitter current to flow.[5] On the other hand, in the FET–IGBT comparison above, the FGA60N65SMD IGBT happens to have a rise/fall time of around 50 ns, while the IRF1405 MOSFET has a rise/fall time of 190/110 ns, respectively.

5.6.1 Driver Circuitry

In both power MOSFET and IGBT switching circuits, the transistor is only thermally well-behaved if the device is fully `OFF` (zero current, large voltage) or fully `ON` (large current, small resistance/voltage). In between, the transistor may drop a substantial voltage while conducting a large current, which can be a disaster if it carries on for long. For this reason, it is a huge advantage to switch between the `ON` and `OFF` states as quickly as possible. However, the gate acts as a capacitive load to the input (5.5 nF for the IRF1405, 2.9 nF for the FGA60N65SMD, 4.6 nF for the IXYN80N90C3H1), and to change the gate voltage rapidly requires a large, transient current. Thus, such high-power devices typically require special gate-driver circuitry, typically consisting of a pair of smaller FETs or BJTs acting as a fast push-pull buffer. These parts are conveniently packaged in some integrated gate-driver IC's; one example with a relatively large current capability is the UCC37321/UCC37322 by TI,[6] which can drive a 9-A peak current, with the output switching between ground and a maximum power supply voltage of +15 V (the UCC37321 inverts the input signal in the sense of a logic inversion, as in Section 9.3.1, while the UCC37322 performs no inversion). An driver example driving up to 340 A (pulsed) current through a load at high voltage via a power IGBT is shown below.



The driver takes a TTL-compatible input (see Section 11.4), but the `ON` level can also be up to +15 V. The driver also has an `ENABLE` input, which forces the output `LOW`, close to 0 V (this is also true on the inverting UCC37321), independent of the input voltage.

A number of other components ensure proper operation and protection of the switching transistor. First, the 5-Ω gate resistor limits the gate current to 15 V/5 Ω = 3 A in the worst case of a short to ground,

[4] `http://www.pwrx.com/pwrx/docs/cm600ha_24a.pdf`

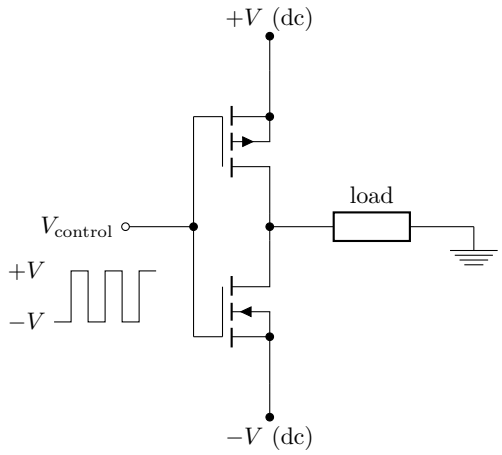
[5] International Rectifier, *Application Note AN-983: IGBT Characteristics*, available online at `http://www.infineon.com/dgdl/an-983.pdf?fileId=5546d462533600a40153559f8d921224`

[6] `http://www.ti.com/lit/ds/symlink/ucc27322.pdf`

ignoring the effects of capacitance, etc. The data sheet implies a minimum value of 2 Ω for this resistance, and gives specifications for up to 15 Ω. This resistance limits the current into the gate capacitance, easing demands on the driver and helping to control against voltage overshoot. The 1N5819 Schottky diodes on the driver output clamp the output voltage to the power-supply range, preventing output overshoot; the Schottky diodes are fast and have a (relatively) small forward voltage. There is also a transient-voltage suppressor (TVS) on the IGBT gate, which is basically two back-to-back Zener diodes. The P6KE20CA TVS break down if the voltage exceeds 20 V in either direction (and is rated for 600 W), protecting agains gate overvoltages (which can come from capacitive coupling via the IGBT as well as from the driver). Finally, as with many IGBTs, the IXYN80N90C3H1 incorporates an internal reverse C–E diode, so that it only stands off voltage in one direction—the device is not designed to stand off high reverse voltages, so the diode offers protection in the event of (intentional or unintentional) reverse transient voltages. However, it is good practice to include another (unidirectional) TVS to protect against reverse and overvoltage transients. The 1.5KE400A TVS acts as a fast Zener diode that breaks down at 400 V and can handle 1.5 kW.

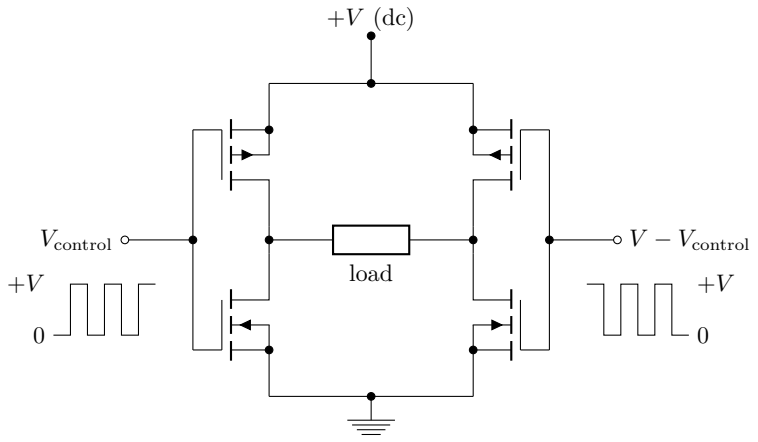
5.6.2 Inverter Circuits

One important application of power MOSFET and IGBT switching circuits is the **inverter**, which is used to convert dc to ac signals. An example of a **half-bridge inverter** is shown below, where a push–pull pair of MOSFETs switches one side of a load between $\pm V$, giving a square-wave ac signal of amplitude V.



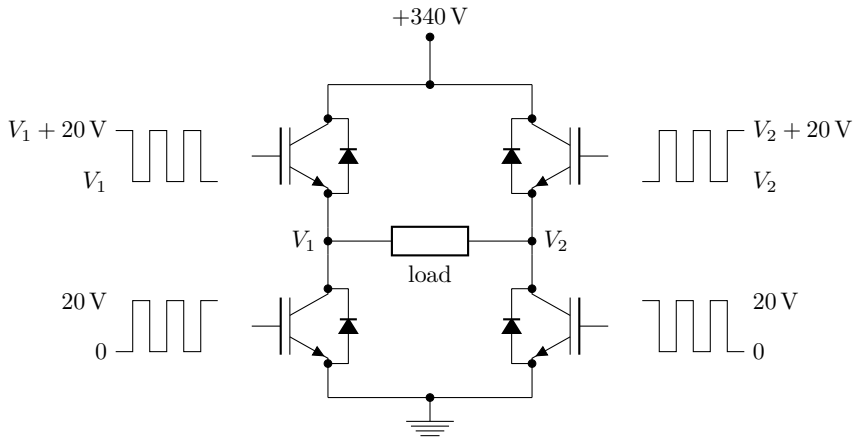
Circuits of this type are particularly common in battery-powered applications. For example, an uninterruptible power supply (UPS) creates an ac waveform from dc battery power, after a dc–dc converter boosts the voltage to 120 V (for U.S. supplies). Cheaper UPSs make a crude approximation of a sine wave by switching to +120 V for a quarter-cycle and −120 V for a quarter-cycle, and to `OFF` for the other quarter cycles in between; this gives both the correct peak and rms voltages, compared to the sine wave. More sophisticated UPSs use rapid switching, pulse-width modulation, and filtering to create a clean sine wave. Other important applications of inverter circuits is in creating ac power from solar cells, motor-driving circuits in battery-powered electric cars, induction "burners" in kitchen stoves, and arc welders. Switching power supplies also use inverters to convert rectified mains voltage into a high-frequency (tens of kHz) signal that can be changed to other dc voltages by a transformer and another rectifier and filter circuit; because of the high frequency, the transformer can be compact and inexpensive compared to traditional 60-Hz transformers. Compared to linear power supplies, they are noisier but cheaper and more compact, and are preferred, for example, in computer power supplies requiring high current.

A circuit that makes better use of the available voltage, at the expense of more parts and a possibly ungrounded load is the **full bridge** or **H-bridge** (after the shape made by the switching transistors and the load), as shown below.

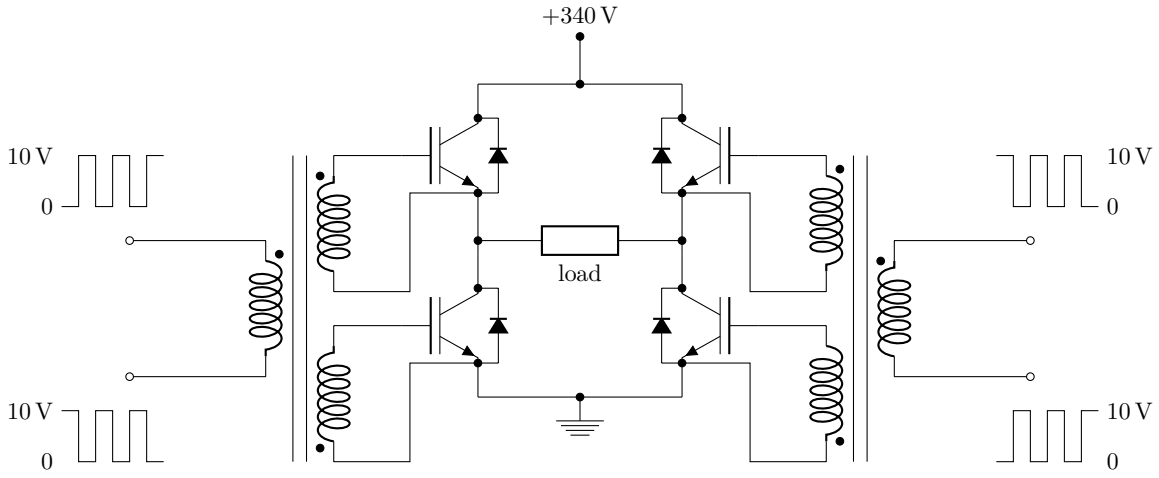


This circuit is basically two MOSFET half-bridges switched out of phase. As in the half-bridge, the control voltage swings between the power-supply rails; the pairing of complementary n-channel and p-channel devices makes the control relatively simple. In each state of V_{control}, a diagonally opposed pair of transistors conducts, while the other pair is open. The result is that the voltage across the load alternates between $\pm V$, but the $-V$ power supply is no longer necessary, as in the half-bridge example above.

For high voltages and high currents, again, IGBTs are preferable. However, IGBTs are generally not available in complementary pairs, which makes the gate control considerably more complicated. An example H-bridge is shown below.



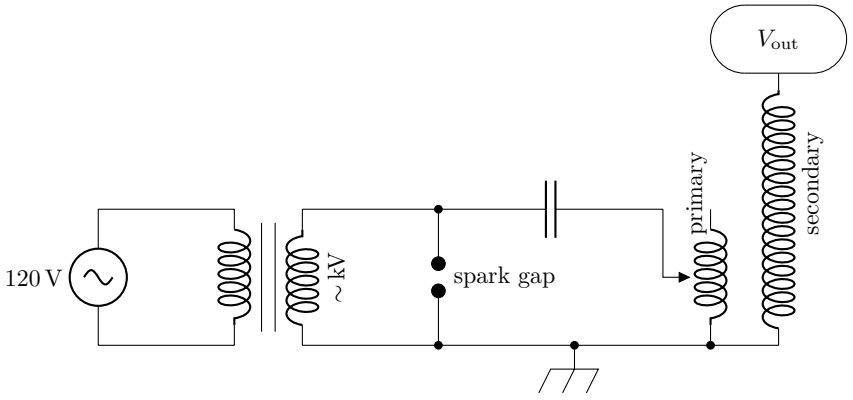
The voltages shown are appropriate for the IXYN80N90C3H1 IGBT, for example. Note that the IGBT gate voltage is always measured with respect to the collector. The lower two IGBT's can thus be switched normally, except the control voltage need not swing all the way to the upper supply rail. The tricky part, however, is that the *upper* two IGBTs require similar control voltages, but referenced to their emitter voltages, which change according to the phase of the inverter. Thus, the control circuitry for these two IGBTs must float along with the appropriate load voltage. One solution to this is to use special driver ICs with optocouplers that isolate the input and allow the output voltage to float with respect to the input. Another solution is to couple the gates via gate-drive transformers, as illustrated below.



In this case, two gate driver ICs drive the ends of the primary coil of a transformer with out-of phase signals, making the primary voltage oscillate between $\pm 20\,\mathrm{V}$. The same drivers can drive the primaries of both transformers, but then the primaries should be connected in antiparallel. The two antiparallel secondaries (with a 1:1 primary:secondary ratio here) drive the gates, and there is no problem floating the gate voltages. Note that this only works for relatively rapid switching and a symmetric drive waveform, as the transformers cannot support a dc signal.

5.6.2.1 Inverter-Based Tesla Coil

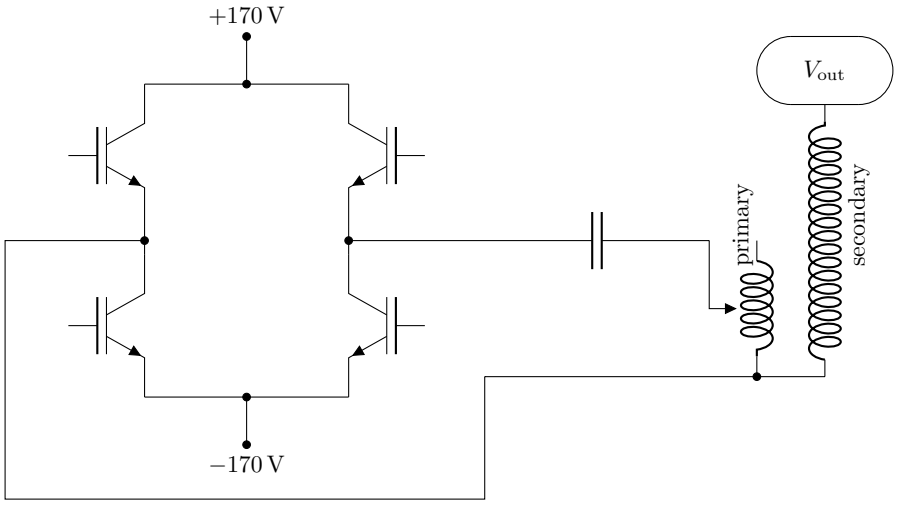
Returning to the Tesla-coil example of Section 2.7.1, suppose that we redraw it from the version we studied before as shown below.



Again, the LC tank circuit resonates at the same frequency as the secondary, and it acts as a harmonic oscillator. The tank circuit is ultimately powered by the 60-Hz transformer output, but because the resonant frequency is much higher ($\sim 100\,\mathrm{kHz}$), we need a way to "whack" the tank circuit to get it oscillating. This is a useful way to look at one function of the spark gap: it holds off the power-supply voltage until it builds up, and then it suddenly lets it into the LC circuit, exciting it with a step voltage.

A more recent variation on the classic Tesla coil involves replacing the spark gap with an inverter circuit.[7]

[7] See, for example, `http://www.stevehv.4hv.org`.

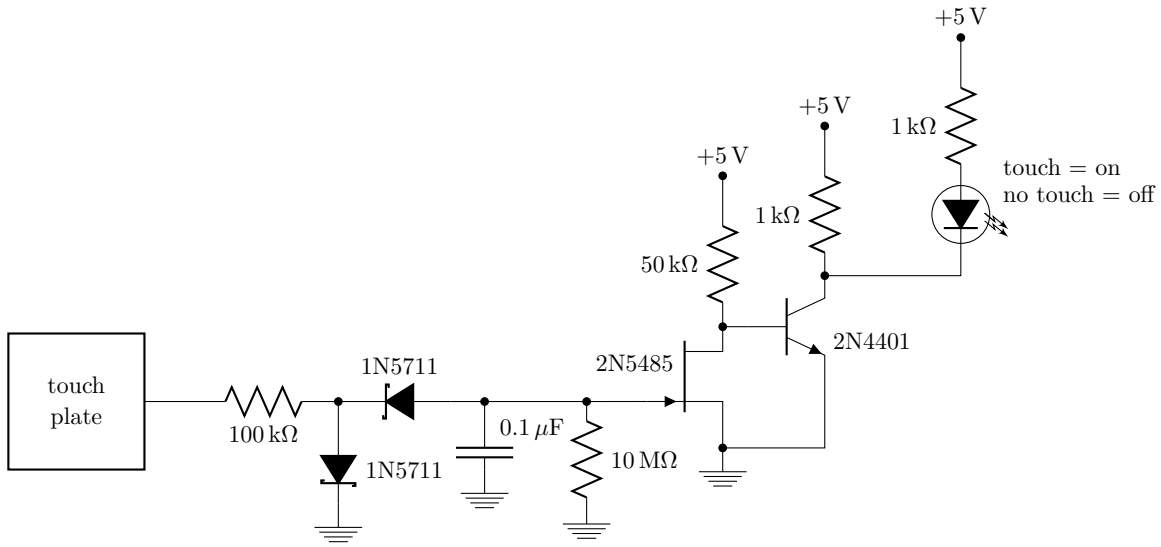


In this example, a dc power supply of ±170 V (derived by rectifying and filtering 120-V mains voltage) is switched by a full-wave bridge to give the same, step-function excitation to the tank circuit. This design can be more consistent in its performance, because it does not rely on a discharge to excite the resonator. However, the overall circuitry is generally much more complicated, involving the driver and protection circuitry discussed above, as well as yet more control circuitry. For example, to not overload the IGBTs, the bridge should only switch when the current in the tank circuit crosses through zero; otherwise, the sudden change in current would lead to a large inductive-kick voltage, which can strain the IGBTs. So a current detector is needed to feed back to the inverter inputs so that switching occurs at the correct time. Inverter-type coils are also typically modulated, being active only a small fraction of the time, so that the high currents involved to not thermally overload the IGBTs. This requires more control circuity which must also turn off the inverter bridge only at the proper times.

5.7 Circuit Practice

5.7.1 Touch Switch

Starting in the 1980's, table lamps with a "touch base" became popular. The idea is, when you touch a metallic base, the lamp toggles between on and off (or between off, on, and two dim states). The key to the operation of this circuit is that the human body acts as an efficient antenna for 60-Hz radiation, which can be coupled into a metallic plate (or even a "dangling" wire) and into a circuit. The circuit below shows how the coupled radiation can be rectified and used as a switch.



- How does this circuit work?
- What is the function of the 10-MΩ resistor?
- Note that Schottky diodes have a small forward voltage (about 0.2 V), to make the circuit more sensitive.
- How would you have to modify the circuit to use a MOSFET in place of the JFET, and can you think of an advantage of this circuit over that one?
- To add in some digital electronics, what can you add to the circuit to toggle the LED between `ON` and `OFF` on each touch?
- How should this circuit be modified to control a 120-V incandescent lamp?

Solution. When nobody is touching the plate, there is essentially no signal at the input; then the FET is `ON` (the default state), which means the npn base is at low voltage. Thus, the npn transistor is `OFF`, so the connection to the LED is pulled to +5 V, so the LED is off. (Strictly, the 1-kΩ resistor at the collector is not necessary, but it allows the circuit to work if the LED is replaced by something else, such as a relay, discussed below.

When touching the plate, 60-Hz radiation is coupled in, and rectified by the diode network (Schottky diodes with small forward voltages) to pass only the negative parts of the signal. The first diode (to ground) discards the positive half of the signal, which shouldn't make it through the second diode anyway; so it seems like this first diode might not really be necessary. However, this first diode gives a path to ground for the positive part of the input signal (so the diode network is basically a charge pump, transferring charge from the capacitor to ground via the diodes). This process charges the capacitor to negative voltage; when this

crosses the threshold (−4 V max for this JFET), the JFET turns `OFF`, which turns `ON` the npn transistor, and lights the LED.

When the touching stops, the 10-MΩ resistor slowly (1-s $1/e$ time) discharges the capacitor to turn the JFET back `ON`. Empirically, the LED turns off in a fraction of a second with a brief touch of the antenna, but can take about 2 seconds to turn off with a long tough of the antenna.

To use a MOSFET, the direction of the diode would have to be reversed, because we'd want to keep the positive parts of the signal to change the MOSFET state. This would also reverse the sense of the LED being on/off, so the LED could be placed between the collector and ground to restore previous operation. Note that MOSFETs are more sensitive to destruction by static discharge and overvoltage. While the 100-kΩ input resistor should help in this regard, it would be safer to use a clamping diode at the gate to make sure the MOSFET is protected. Overall, the JFET makes for a simpler design here.

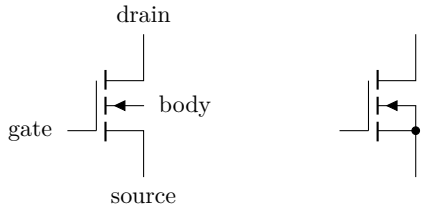
To get the LED to toggle, a divide-by-2 counter (flip-flop) would help here.

To drive a heavy load like a light bulb, the output of this circuit should drive a relay (or even a small relay that drives a large relay, which in turn drives the light bulp). It could also drive a solid-state version of a relay, like a power-MOSFET switch.

5.8 Exercises

Problem 5.1

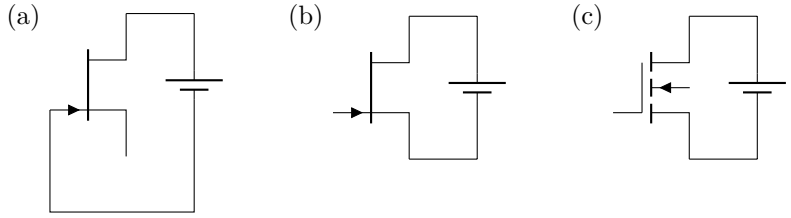
Consider an n-channel (enhancement-mode) MOSFET. The standard connections are shown on the left in the diagram below. Recall that if the G–S voltage is zero, the MOSFET is in the "pinched-off" state, so no D–S current flows in either direction.



Many MOSFETs have the body connected internally to the source (often to fit the MOSFET into a cheaper, 3-pin package), as shown to the right in the diagram. In this case, when $V_{GS} = 0$, the MOSFET will block current *in one direction only*, while conducting freely in the other direction. Which direction is which? Explain *why* in terms of the underlying n-type and p-type semiconductor regions.

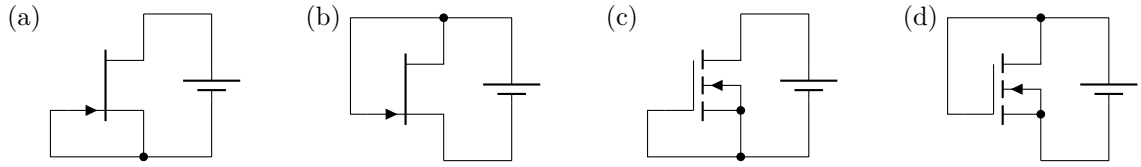
Problem 5.2

In each circuit, does current flow due to the applied EMF? (Ignore leakage currents, e.g., in a reverse-biased diode, and assume the applied EMF is a few volts.) **Briefly** explain your answer.



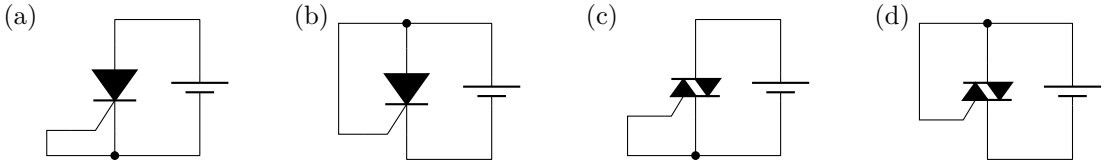
Problem 5.3

In each circuit, does current flow due to the applied EMF? (Ignore leakage currents, e.g., in a reverse-biased diode, and assume the applied EMF is a few volts.) **Briefly** explain your answer.



Problem 5.4

In each circuit, does current flow due to the applied EMF? Assume the applied EMF is enough to cause a forward-biased p–n junction to conduct, but **not** enough to break down a reverse-biased p–n junction. Also ignore any leakage currents, and assume the device was `OFF` when the battery was connected. **Briefly** explain your answer, and indicate the paths along which current flows.



Chapter 6

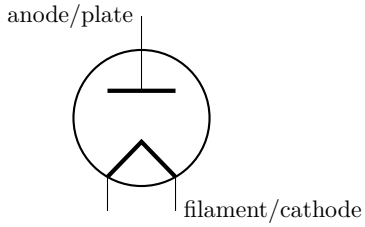
Vacuum Tubes

Nowadays we take it for granted that there can be billions of transistors in a modern computer. But before transistors there were vacuum tubes, which are much larger, hotter, higher-voltage, and more power-consuming than their silicon successors. Despite the low cost and ubiquity of transistors, vacuum tubes still have a few important applications, from audio amplifiers to circuits with very high power (e.g., for radio-frequency and microwave applications).

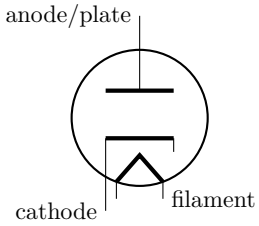
This chapter is located just after FETs because they fit in logically here. However, vacuum tubes remain something of a niche subject. They're worth studying for fun and more circuit practice, but ultimately the op-amps in the next chapter are much more useful: make sure to learn about them first and come back to tubes later if you're still interested.

6.1 Vacuum Diodes

The simplest vacuum tube is the **vacuum diode**, so named because of the two fundamental elements, the anode and the cathode. The main function of the cathode is to act as a source of emitted electrons. There are two basic types. The simpler **filamentary cathode** or **directly heated cathode** is simply a filament of wire, heated to high temperature (typically dull glowing red, typically 700°C). The wire is coated so that electrons "boil" easily from the surface at operating temperature. The anode, often called the **plate** is a conductor that attracts or repels the electrons for a positive or negative voltage, respectively, relative to the cathode. Current only flows through the tube in the positive-voltage case where the electrons are drawn to the anode. The symbol for the diode with directly heated cathode is shown below.



The other type of cathode is the **heated cathode indirectly heated cathode**. In this case the cathode is physically and electrically separate from the heating filament, but the cathode still becomes hot and acts as an electron source. In this arrangement the cathode acts as a shield for any ac fields produced by the filament, which is typically heated with a low-voltage ac current. This is an advantage in amplifier vacuum tubes to avoid a filament-induced "hum" in the output, and the vast majority of more sophisticated tubes below have indirectly heated cathodes. However, vacuum diodes in power-supply circuits don't benefit greatly in this respect, so filamentary cathodes are common. The symbol for a diode with an indirectly heated cathode is shown below.



6.1.1 Child–Langmuir Law

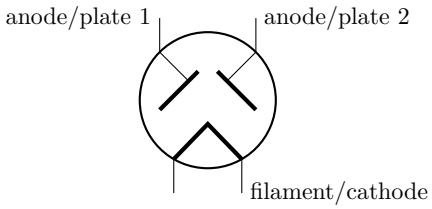
For either cathode type, as the cathode temperature increases, the density of emitted electrons increases, leading to an increased current in forward conduction mode. However this increase does not continue indefinitely: the electrons themselves will modify the electric field between the cathode and anode, and the net effect is that the current is limited to

$$I = \frac{4\epsilon_0 A}{9}\sqrt{\frac{2e}{m_\mathrm{e}}}\frac{V^{3/2}}{d^2}. \tag{6.1}$$

This expression, called the **Child–Langmuir Law**,[1] assumes parallel, identical plates for the anode and cathode, where A is the anode (cathode) area, e is the electron-charge magnitude, m_e is the electron mass, V is the diode forward voltage, and d is the anode–cathode distance. This scaling of $I \propto V^{3/2}$ holds in other geometries, but the proportionality factor is geometry-dependent.[2] This law can act as a rough guide to the voltage–current characteristics of a vacuum tube, although it is not accurate enough for quantitative design purposes. However, it is clear from this law that the turn-on of a vacuum tube with voltage is much less "sudden" than the semiconductor diode, which turns on exponentially with forward voltage.

6.1.2 Vacuum Full-Wave Rectifier

In semiconductor full-wave rectifier circuits, it is typical to use a full-wave bridge consisting of four diodes (see Section 3.6.2). However, vacuum diodes are more complicated and expensive, so such an arrangement is rare. However, to economize on the number of vacuum tubes, a common variation on the vacuum diode is the **vacuum full-wave rectifier**. This is the same as a vacuum diode, but with two anodes, forming a bridge of two diodes. The schematic symbol is shown below.



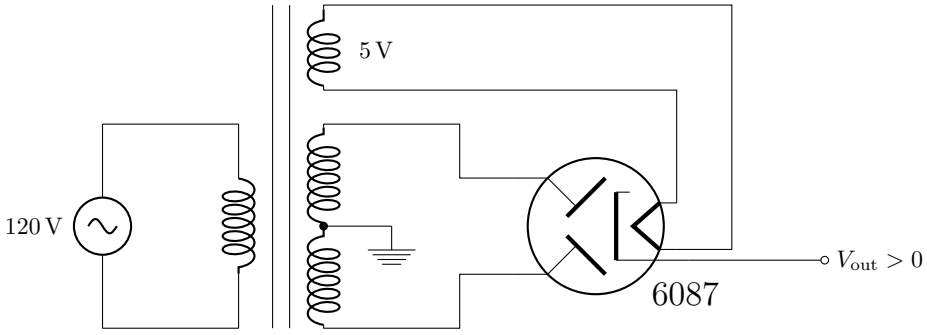
A common, low-power rectifier tube is the 5Y3, a directly heated full-wave rectifier. The part number in this case has a typical numbering that is happily informative: the "5" says the filament operates at 5 V, and the trailing "3" says that there are three elements (two anodes and one cathode). The "Y" is a symbol to differentiate the tube. The part is often listed as, for example, 5Y3GT, where the "GT" indicates a "glass tube" envelope. The 5Y3 operates with 2 A of filament current, can stand off 1400 V of reverse voltage, and can handle 400 mA max forward current (in steady state, rising to 2.5 A max transient current). In typical

[1] C. D. Child, "Discharge From Hot CaO," *Physical Review (Series I)*, **32**, 492 (1911) (doi: 10.1103/PhysRevSeriesI.32.492); Irving Langmuir, "The Effect of Space Charge and Residual Gases on Thermionic Currents in High Vacuum," *Physical Review* **2**, 450 (1913) (doi: 10.1103/PhysRev.2.450).

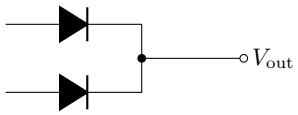
[2] For example, for long. coaxial cylinders, the relation becomes $I = (8\pi\epsilon_0 L/9)\sqrt{2e/m_\mathrm{e}}V^{3/2}/r\beta^2$, where L is the cylinder length, V is the potential at the radius r, and β is a dimensionless factor of order unity. See Irving Langmuir and Katharine B. Blodgett, "Currents Limited by Space Charge between Coaxial Cylinders," *Physical Review* **22**, 347 (1923) (doi: 10.1103/PhysRev.22.347).

operation, it would operate at a 360-V (dc) plate voltage, and 125 mA of output current, with a forward voltage drop of 50 V (well above the ∼0.7 V for a semiconductor diode!). To characterize the reverse leakage, the tube is specified with a reverse resistance of around 10 MΩ.

A typical full-wave rectifier circuit using a 6087 rectifier tube (essentially the same as the 5Y3, but with indirect heat) appears below.



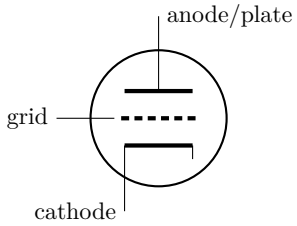
A transformer steps up the mains voltage to several hundred volts (a typical value is 320 V rms, which translates to 453 V peak, not accounting for voltage dropped by the internal resistance of the secondary winding). Again, the tube drops about 50 V, and accounting for the secondary output impedance, the output voltage drops to around 320 V (dc peak). Note that the transformer output is center-tapped, as needed to make a full-wave rectifier with only two diodes. A four-diode full-wave rectifier would only need half the secondary winding, but the extra complexity of the transformer is an acceptable tradeoff to reduce the tube count. The vacuum rectifier again acts as a diode pair, as illustrated below.



The heater in the tube is driven by a separate, low-voltage winding on the same transformer.

6.2 Vacuum Triodes

The next tube in the hierarchy, the simplest tube intended for use as an amplifier, is the **vacuum triode**. The name of course indicates that there is now a third element, the **grid** or **control grid**. The grid is something like a fine-mesh electrode between the anode and cathode, but is often made from a fine wire wrapped around supporting posts. The schematic symbol for a triode appears below.



The idea is that the grid can modulate the plate current by attracting or repelling electrons, in the same way as the anode voltage controls current. The grid area is small (i.e., the wire mesh is mostly open), so it is intended only to modulate the plate current, not to conduct any of the electrons itself. In typical operation, the plate–cathode voltage $V_{PC} > 0$ is fixed, while the grid–cathode voltage $V_{GC} < 0$ and is relatively small in magnitude. Thus the grid tends to oppose the action of the anode, and a more negative V_{GC} reduces the plate current. In this sense, the triode grid is analogous to the gate of an n-channel (depletion-mode) JFET, where an increasingly negative voltage pinches off the drain–source current (see Section 5.1). However, the

scale of the voltages is different, and, unlike for the JFET, the "pinch-off" voltage for the triode grid depends on the plate voltage.

In the case of a diode, the Child–Langmuir Law stated that the plate–cathode current and voltage are related by $I_{\mathrm{PC}} \propto V_{\mathrm{PC}}^{3/2}$, where the proportionality constant is geometry-dependent. We can adapt this law to the triode by writing

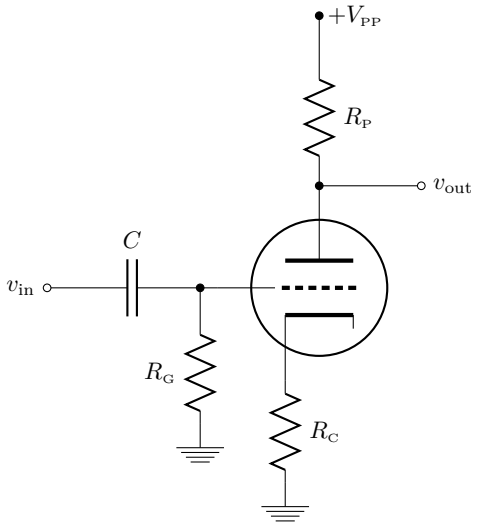
$$I_{\mathrm{PC}} \propto (\mu V_{\mathrm{GC}} + V_{\mathrm{PC}})^{3/2} \qquad (\mu V_{\mathrm{GC}} + V_{\mathrm{PC}} > 0), \tag{6.2}$$

where μ is the triode **magnification factor**, and represents the ratio of the two proportionality constants for the two geometries (grid–cathode to plate–cathode). The magnification factor typically falls in the range of 10 to 100 or more. In this simple model, the current is zero if the tube is "reverse-biased" ($\mu V_{\mathrm{GC}} + V_{\mathrm{PC}} < 0$). Again, the above triode law only acts as a rough guide to triode operation; triode circuits are usually analyzed graphically or on the computer for quantitatively accurate results.

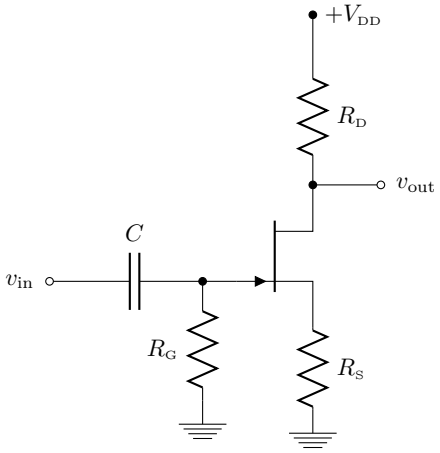
An example of a popular triode tube is the 12AX7 dual triode (i.e., 2 triodes in one package; "12" = 12.6-V filament, center-tapped to also work with 6.3 V; "7" = 7 elements: 2 anodes, 2 grids, 2 cathodes, 1 heater). This tube is especially popular in audio preamplifiers. It typically operates with a plate voltage V_{PC} of 100 to 250 V. The typical grid voltage V_{GC} is -1 to -2 V; the magnification factor is $\mu = 100$; the grid transconductance g_{m} is in the range of 1250 to 1600 $\mu\mho$; and the typical plate current is 0.5 to 1.2 mA.

6.2.1 Triode Voltage Ampifier (Common-Cathode Amplifier)

A typical circuit for using a triode as an amplifier, the **common-cathode amplifier**, is shown below.



Again, due to the analogy with the JFET, the tube circuit is very similar to the JFET (common-source) voltage amplifier shown below, which we analyzed before (Section 5.4.3).



In both cases, the grid/gate is biased at 0 V, and the cathode/source resistor functions to elevate the cathode/source voltage so the grid gate is biased slightly negative with respect to the cathode/source. In the vacuum-tube case, this is called a **cathode-biased** amplifier.

The common-cathode amplifier will generally also drive a load, which we can regard as having a Thévenin resistance R_L. The net effect is that R_P should be replaced by $R_P \| R_L$, and the supply voltage V_{PP} should be replaced by the appropriate linear combination of the supply voltage and the Thévenin voltage of the load. By using R_P and V_{PP}, the following analysis also handles a loaded amplifier if these modifications are kept in mind.

6.2.1.1 DC Bias

As an example of tube-amplifier analysis, we will work through the analysis of the triode amplifier here. The plate power supply $+V_{PP}$, powers the arrangement, supplies plate–cathode current, and as a result it again sets up the bias voltage for the grid. In traditional lingo, the plate power supply is called the "B+ voltage," in reference to the three elements A, B, and C of the triode ("A" being the cathode and "C" being the grid), and the A, B, and C batteries that were used to power the early circuits.

To begin, the dc current is determined by the power-supply voltage, which must be dropped across the plate and cathode resistors as well as the tube itself along the plate–cathode path:

$$V_{PP} - V_{PC} = I_{PC}(R_P + R_C). \tag{6.3}$$

This is called a "load-line equation," because when written as a relation for the the plate current I_{PC} as a function of the plate–cathode voltage V_{PC}, this becomes the equation for a line with slope $-1/(R_P + R_C)$ and intercept $V_{PP}/(R_P + R_C)$:

$$I_{PC} = \frac{V_{PP} - V_{PC}}{R_P + R_C}. \tag{6.4}$$

This current then determines the grid bias point via

$$V_{GC} = -V_C = -I_{PC}R_C. \tag{6.5}$$

Finally, the tube itself has a response current $I_{PC}(V_{GC}, V_{PC})$ to the grid and plate voltages. In general, this response function does not have a simple form, so the design is often done graphically on a plot of I_{PC} vs. V_{PC} for various values of V_{GC}. The load line and grid-bias relations are superimposed on this plot, and the intersection of the two determines the operating point. However, we can do this using the Child–Langmuir Law (6.2) to illustrate the process algebraically (at the cost of accuracy):

$$I_{PC} = a(\mu V_{GC} + V_{PC})^{3/2}. \tag{6.6}$$

Using Eqs. (6.3) and (6.5) to eliminate the voltages, the result is

$$I_{PC} = a\Big[V_{PP} - \big[(\mu + 1)R_C + R_P\big]I_{PC}\Big]^{3/2}. \tag{6.7}$$

This is effectively a cubic equation in $I_{\rm PC}^{1/3}$, and thus has an analytic (but messy) solution. The point is that it can in principle be solved for $I_{\rm PC}$, given the resistances and the power-supply voltage. Once this dc current is known, this fixes the grid bias via Eq. (6.5).

6.2.1.2 Naïve AC Analysis

Once the grid bias is known, the grid (ac) transconductance

$$g_{\rm m} := \frac{\partial I_{\rm PC}}{\partial V_{\rm GC}} \qquad (6.8)$$

(tube transconductance: definition)

is then known. In the Child–Langmuir model the transconductance is

$$g_{\rm m} = \frac{\partial I_{\rm PC}}{\partial V_{\rm GC}} = \frac{3a\mu}{2}(\mu V_{\rm GC} + V_{\rm PC})^{1/2}. \qquad (6.9)$$

Note that even though the Child–Langmuir model is a rough approximation, the derivative still serves to define the transconductance, and the concept hold in general (though with a different quantitative dependence on the plate and grid voltages).

Now we will proceed with an analysis that traces along the lines of the JFET voltage amplifier. However, it yields expressions that in general do not work well for triode amplifiers (although for pentodes below this analysis works well). Afterwards, we will show how to improve it to handle triode amplifiers.

Given a small input ac voltage $v_{\rm in}$ (on top of a dc bias, as in the previous section), assuming a high enough frequency that the input-capacitor impedance is negligibly small, the input voltage causes an ac plate current (i.e., a small change in $I_{\rm PC}$)

$$i_{\rm PC} = g_{\rm m}(v_{\rm G} - v_{\rm C}) = g_{\rm m}(v_{\rm in} - i_{\rm PC}R_{\rm C}). \qquad (6.10)$$

Solving for $i_{\rm PC}$,

$$i_{\rm PC} = \frac{g_{\rm m}}{(1 + g_{\rm m}R_{\rm C})}v_{\rm in}. \qquad (6.11)$$

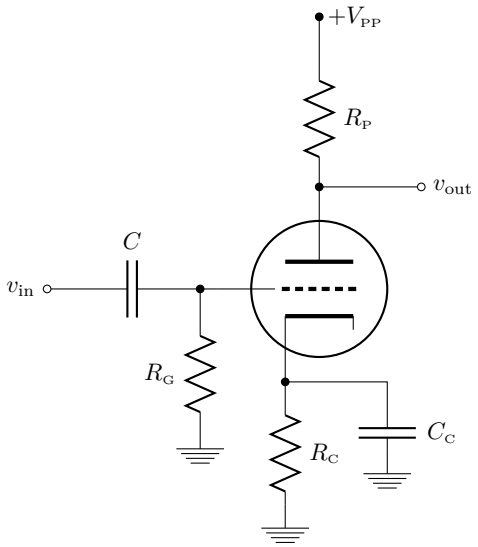
Then this current determines the ac output via

$$v_{\rm out} = -i_{\rm PC}R_{\rm P} = -\frac{g_{\rm m}R_{\rm P}}{(1 + g_{\rm m}R_{\rm C})}v_{\rm in}, \qquad (6.12)$$

for an ac voltage gain

$$G = -\frac{g_{\rm m}R_{\rm P}}{(1 + g_{\rm m}R_{\rm C})}. \qquad \text{(naïve!)} \qquad (6.13)$$

Note that in Eq. (6.10) leading up to this, the effect of the cathode resistor was to take away some of the gain, because an increase $v_{\rm in}$ causes increasing $i_{\rm PC}$, but this somewhat counteracts the input voltage via an increase in $v_{\rm C}$. If desired, this effect may be counteracted by including a bypass capacitor across $R_{\rm C}$, which removes its effect for ac signals, as shown below.



In this case, the gain is simply

$$G = -g_{\mathrm{m}} R_{\mathrm{P}}. \qquad \text{(still naïve!)} \tag{6.14}$$

As in analogous transistor circuits, the input impedance is controlled mainly by R_{G}, and the output impedance is set mainly by R_{P}.

6.2.1.3 Proper AC Analysis: Plate Resistance

The important complication that we left out of the above analysis is the **plate resistance**: the plate current depends not only on the gate voltage V_{GC}, but also on the plate voltage V_{PC}. In the JFET-amplifier analysis of Section 5.1, we explicitly assumed the saturation regime, where the drain–source current is (approximately) independent of the drain–source voltage. But this same assumption does not carry over in the same way in many tube-amplifier circuits.

The plate resistance r_{p} is defined by the current–voltage slope

$$\frac{1}{r_{\mathrm{p}}} := \frac{\partial I_{\mathrm{PC}}}{\partial V_{\mathrm{PC}}}, \tag{6.15}$$
(plate resistance: definition)

in analogy to the intrinsic emitter resistance of the BJT [see Eq. (4.69)]. For a 12AX7 tube, r_{p} is typically in the range of $60 - -80\,\mathrm{k\Omega}$, which is usually significant on the scale of the other resistances in the circuit.

Now, to account explicitly for the ability of either the gate or plate voltage to influence the plate current, we can write a differential change in the current as

$$\begin{aligned} dI_{\mathrm{PC}} &= \frac{\partial I_{\mathrm{PC}}}{\partial V_{\mathrm{PC}}}\, dV_{\mathrm{PC}} + \frac{\partial I_{\mathrm{PC}}}{\partial V_{\mathrm{GC}}}\, dV_{\mathrm{GC}} \\ &= \frac{dV_{\mathrm{PC}}}{r_{\mathrm{p}}} + g_{\mathrm{m}}\, dV_{\mathrm{GC}}, \end{aligned} \tag{6.16}$$

after using Eqs. (6.8) and (6.15). Suppose that I_{PC} is held constant; then the right-hand side of the above equation vanishes, leading to the relation

$$g_{\mathrm{m}} r_{\mathrm{p}} = -\frac{\partial V_{\mathrm{PC}}}{\partial V_{\mathrm{GC}}}. \tag{6.17}$$

To evaluate the derivative, going back to the triode Child–Langmuir relation (6.2), at constant I_{PC} this relation leads to

$$\mu V_{\mathrm{GC}} + V_{\mathrm{PC}} = \text{const.} \tag{6.18}$$

if V_{GC} and V_{PC} change by infinitesimal amount dV_{GC} and dV_{PC}, respectively, this relation implies that they must cancel, and thus

$$\mu = -\frac{\partial V_{PC}}{\partial V_{GC}}. \tag{6.19}$$

This of course holds for more general models where the current depends on the combination $\mu V_{GC} + V_{PC}$, but more fundamentally, we can view this partial derivative as the definition of the magnification factor, as it expresses the relative effect of the two voltages on the current via $\mu = -(\partial I_{PC}/\partial V_{GC})/(\partial I_{PC}/\partial V_{GC})$. Combining this relation with Eq. (6.17), this leads to the relation

$$\mu = g_m r_p \tag{6.20}$$

(small-signal-parameter relation)

connecting the various small-signal tube parameters

With the above development in hand, we can return to the ac gain of the common-cathode amplifier on p. 126. Using Eqs. (6.16) in terms of biased ac signals, we have

$$\begin{aligned} i_{PC} &= g_m v_{GC} + \frac{v_{PC}}{r_p} \\ &= g_m(v_{in} - v_C) + \frac{v_{PC}}{r_p}. \end{aligned} \tag{6.21}$$

To evaluate v_{PC}, note that the plate–cathode voltage drops if I_{PC} rises, because voltage drops across R_P and R_C increase:

$$v_{PC} = -i_{PC}(R_P + R_C). \tag{6.22}$$

Putting this and $v_C = i_{PC}R_C$ into Eq. (6.21) gives

$$i_{PC} = g_m(v_{in} - i_{PC}R_C) - \frac{R_P + R_C}{r_p} i_{PC}, \tag{6.23}$$

or after solving for i_{PC}, we find

$$i_{PC} = \frac{g_m r_p v_{in}}{r_p + (g_m r_p + 1)R_C + R_P} = \frac{\mu v_{in}}{r_p + (\mu + 1)R_C + R_P} \tag{6.24}$$

after applying Eq. (6.20). Then using $v_{out} = -i_{PC}R_P$ and $G = v_{out}/v_{in}$, the voltage gain becomes

$$G = -\frac{\mu R_P}{r_p + R_P + (\mu + 1)R_C}.$$

(common-cathode amplifier voltage gain) (6.25)

For the amplifier on p. 128 where R_C is bypassed by a capacitor, we can set $R_C = 0$ in the gain expression, which reduces to

$$G = -\frac{\mu R_P}{r_p + R_P}.$$

(common-cathode amplifier voltage gain, with cathode bypass) (6.26)

It is only in the limit of *large* plate resistance (so I_{PC} is weakly affected by V_{PC}) that the naïve from the previous section follow. Rewriting Eq. (6.25) as

$$G = -\frac{g_m R_P}{1 + R_P/r_p + (g_m r_p + 1)R_C/r_p}, \tag{6.27}$$

we can see that taking the limit $r_p \longrightarrow \infty$ leads to

$$G = -\frac{g_m R_P}{1 + g_m R_C},$$

(common-cathode amplifier voltage gain, infinite plate resistance) (6.28)

in agreement with Eq. (6.13). Bypassing the cathode capacitor then leads to the expression $G = -g_m R_P$, as in Eq. (6.14). This situation with large r_p is more likely to occur in pentode and beam power tubes, as described below, when the plate resistor (and load) are relatively small compared to the plate resistance. As far as obtaining a large gain is concerned, a larger r_p leads to larger gain.

6.2.2 Design Example: 12AX7 Preamplifier

As a specific example, consider a preamplifier stage made from (half of) a 12AX7, with

- $R_{\mathrm{P}} = 100\,\mathrm{k\Omega}$,
- $R_{\mathrm{C}} = 1.5\,\mathrm{k\Omega}$,
- $R_{\mathrm{G}} = 1\,\mathrm{M\Omega}$ (the grid resistor only really matters for the input impedance and to ensure a proper grid bias),
- $V_{\mathrm{PP}} = 330\,\mathrm{V}$, and
- model values of $a = 1.89 \times 10^{-6}\,\mathrm{A/V^{3/2}}$ and $\mu = 100$ for the 12AX7 (see the next section),

a numerical solution of Eq. (6.7) yields

$$I_{\mathrm{PC}} \approx 1.04\,\mathrm{mA}, \tag{6.29}$$

which leads to a grid bias from Eq. (6.5) of

$$V_{\mathrm{GC}} = -1.57\,\mathrm{V}, \tag{6.30}$$

a tube plate–cathode voltage from Eq. (6.3) of

$$V_{\mathrm{PC}} = 224\,V, \tag{6.31}$$

and a grid transconductance from Eq. (6.9) of

$$g_{\mathrm{m}} = 2330\,\mu\mho. \tag{6.32}$$

Note that the transconductance falls above the range quoted for the 12AX7 in the first part of Section 6.2. With $\mu = 100$ by assumption and

$$r_{\mathrm{p}} = \mu/g_{\mathrm{m}} = 43.0\,\mathrm{k\Omega} \tag{6.33}$$

from Eq. (6.20), the amplifier gain is

$$G = -34, \tag{6.34}$$

from Eq. (6.25) [$G = -52$ from the naïve expression (6.13)], or if the cathode resistor is bypassed, the gain is

$$G = -70 \tag{6.35}$$

[$G = 233$ from the naïve expression (6.14)]. Again, these values should not be taken too seriously, given the simple Child–Langmuir model, although we will see below that they aren't too bad in this example.

6.2.3 Phenomenological Triode Model

It is possible to model the triode behavior more accurately than via the Child–Langmuir model. Koren, for example, proposed the following phenomenological model for a triode:[3]

$$\begin{aligned} V_1 &= \frac{V_{\mathrm{PC}}}{k_{\mathrm{P}}} \log\left\{1 + \exp\left[k_{\mathrm{P}}\left(\mu^{-1} + \frac{V_{\mathrm{GC}}}{\sqrt{k_{\mathrm{VB}} + V_{\mathrm{PC}}^2}}\right)\right]\right\} \\ I_{\mathrm{PC}} &= \frac{V_1^x}{k_{\mathrm{G}}}\big[1 + \mathrm{sgn}(V_1)\big]. \end{aligned} \tag{6.36}$$

Note that when $V_{\mathrm{PC}}^2 \gg k_{\mathrm{VB}}$, $k_{\mathrm{P}}(\mu^{-1} + V_{\mathrm{GC}}/V_{\mathrm{PC}}) \gg 1$, and $x = 3/2$, these equations reduce to the Child–Langmuir Law (6.6) in the form $I_{\mathrm{PC}} = (V_{\mathrm{PC}}/\mu + V_{\mathrm{GC}})^{3/2} 2/k_{\mathrm{G}}$, so that we can identify $a = 2/k_{\mathrm{G}}\mu^{3/2}$.

[3] Norman L. Koren, "Improved vacuum tube models for SPICE simulations. Part 1: Models and example," `http://www.normankoren.com/Audio/Tubemodspice_article.html` (update 2003).

Koren's parameters for the 12AX7 read as follows:

$$\begin{aligned} \mu &= 100 \\ k_{\mathrm{G}} &= 1060\,\mathrm{V}^x/\mathrm{A} \\ k_{\mathrm{P}} &= 600\,\mathrm{V} \\ k_{\mathrm{VB}} &= 300\,\sqrt{\mathrm{V}} \\ x &= 1.4. \end{aligned} \tag{6.37}$$

Then solving Eqs. (6.36) numerically, along with the load-line equation (6.4) and the grid-bias equation (6.5), we find

$$\begin{aligned} I_{\mathrm{PC}} &= 1.07\,\mathrm{mA} \\ V_{\mathrm{PC}} &= 221\,\mathrm{V} \\ V_{\mathrm{GC}} &= -1.61\,\mathrm{V}. \end{aligned} \tag{6.38}$$

To "seed" the root-finding process, the Child–Langmuir results from the previous section may be used as initial guesses. While the previous results are fairly close (they were within a few percent of these values), it is evident that using the better model makes a difference in the design results. Now the transconductance follows from differentiating Eqs. (6.36),

$$g_{\mathrm{m}} = \frac{\partial I_{\mathrm{PC}}}{\partial V_{\mathrm{GC}}} = 1880\,\mu\mho, \tag{6.39}$$

and Eq. (6.20) gives

$$r_{\mathrm{p}} = \frac{\mu}{g_{\mathrm{m}}} = 53.3\,\mathrm{k}\Omega, \tag{6.40}$$

which is somewhat low compared to the RCA specification[4] Eq. (6.25) then leads to a circuit gain of

$$G = -33 \tag{6.41}$$

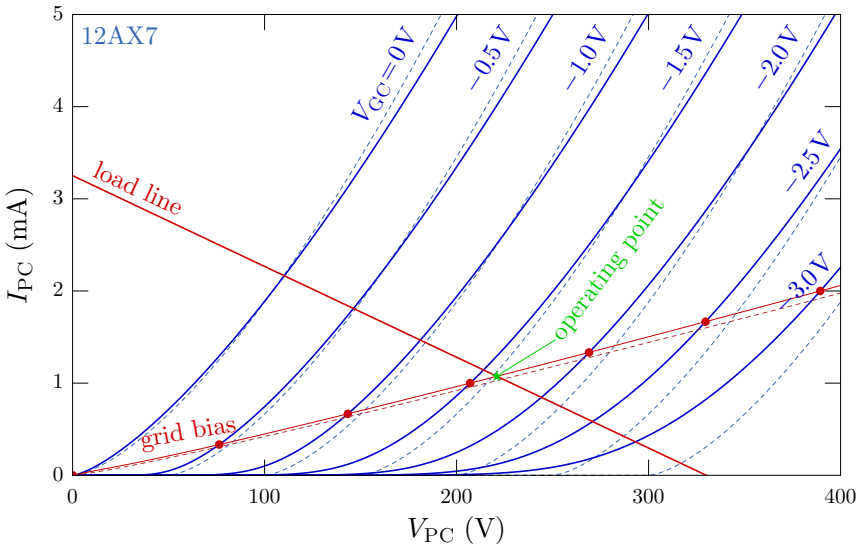
with the feedback from R_{C}, or

$$G = -65 \tag{6.42}$$

with R_{C} bypassed. The Child–Langmuir versions of the gain values were more than 10% from the correct values.

So is this accuracy typical for the Child–Langmuir model? To see this, and to also see how to do this calculation graphically (in case only graphical tube data are available from a data sheet), the plot below shows the plate I_{PC}–V_{PC} characteristics for various gate voltages, shown as the solid curves (from the Koren model). The accompanying dashed curves show the Child–Langmuir predictions by comparison; note that these agree in some regions, but not in others.

[4] Available at `http://drtube.com/datasheets/12ax7-rca1962.pdf`, see the plot on the last page.

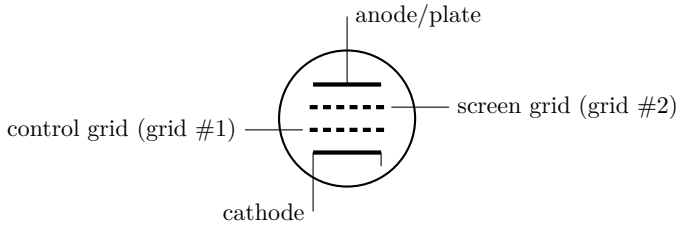


To solve this system, the next step is to draw in the load line (6.3), which is independent of any tube characteristics. Then, draw in the grid-bias curve (6.5) by drawing the point on each curve of constant V_{GC} that satisfies Eq. (6.5) (i.e., by tracing along the curve to the proper current $I_{PC} = -V_{GC}/R_C$). Then the idea is to fill in the curve; while it is not a straight line, a straight line is a reasonably good approximation near any point. The intersection of the load line and the grid-bias curve determine the (dc) operating point of the circuit, and the slope of the grid-bias curve at the operating point gives the transconductance g_m. Alternately, it is possible to more directly obtain the ac amplifier gain, in the case of a bypassed R_C: ac changes in the grid voltage don't affect the bias, so one can follow the load line to find the change in I_{PC} for a change in V_{GC}, and convert to a gain via R_P.

From this diagram it is evident that the circuit happens to operate in the region where the Child–Langmuir model works reasonably well. For tubes in a push–pull circuit, which operate near zero current when the output signal crosses through zero, the simple model makes for a much poorer approximation. Note that the model (6.36) is not perfect either: compared to the RCA data sheet, for example,[5] the $V_{GC} = 0$ curve has a "kink" with the opposite curvature for small V_{PC}. However, this model is probably reasonable for estimates in practical designs.

6.3 Vacuum Tetrodes

While the triode can work well, it has an issue, the **Miller capacitance** between the plate and grid. This is analogous to the Miller effect in bipolar transistors that we mentioned before in Section 4.11.5.3, but of course the vacuum-tube version came first. The Miller capacitance can cause a loss of bandwidth, but even worse, it can cause instability in high-gain circuits, because it acts as a feedback coupling between the high-current output and the sensitive input of an amplifier stage. One solution to this was to introduce a second grid, called the **screen grid** (also called **grid #2**, with the control grid going by the cheeky alternative **grid #1**). The resulting tube is the **tetrode**, in reference to the fourth element. The schematic symbol for the tetrode is shown below.



[5] `http://www.r-type.org/pdfs/12ax7.pdf`

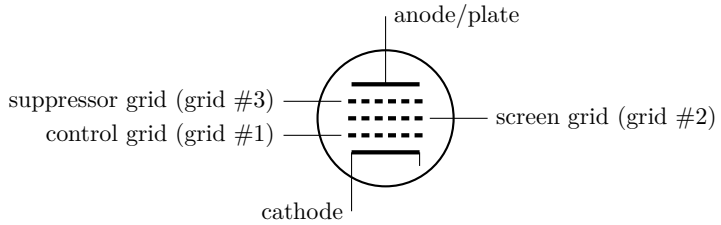
The screen grid sits between the plate and the control grid. The idea is to have a positive, fixed potential of the screen grid, $V_{G_2C} > 0$, but still negative with respect to the plate ($V_{G_2C} < V_{PC}$). Typically this voltage is set via a voltage divider from the plate voltage, and is preferably also bypassed to ground with a capacitor. The control grid is still biased negative ($V_{G_1C} < 0$). Since the screen grid is at fixed potential, it acts as a shield (hence the name "screen," in reference to electrical screening), breaking the capacitive coupling between plate and control grid. The new grid can reduce the control-grid–plate capacitance from several pF to below 0.01 pF.

Since $V_{G_2C} > 0$, the screen grid also attracts electrons from the cathode, which we will see can be a problem. However, this has an advantage, since the screen grid shields somewhat the effect of the anode on the cathode electrons. The net effect is that the plate current I_{PC} is relatively insensitive to the plate voltage V_{PC} (for sufficiently large V_{PC}), which can be a useful property.

6.4 Vacuum Pentodes and Beam Power Tubes

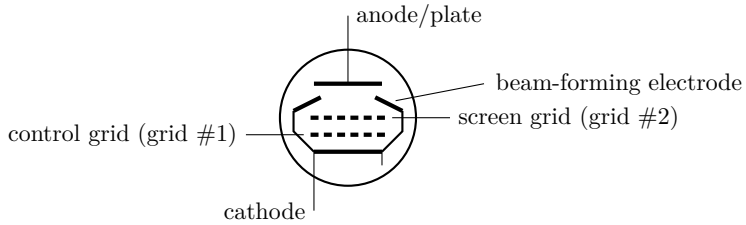
The tetrode, alas, has yet another problem, related to **secondary emission**, which is the emission of electrons from the anode due to cathode electrons striking it. In diodes and triodes, secondary-emission electrons are simply attracted back to the anode, and so don't affect things much (except through space-charge effects). However, in tetrodes, the secondary electrons are attracted to the screen, especially if the screen grid is biased above the plate. The net effect is the reduce the plate current and overall gain of the circuit, and it reduces the useful range of V_{PC}.

The solution is to add yet another grid, the **suppressor grid** or **grid #3**, forming the **pentode**. The suppressor grid site between the screen grid and plate; usually it is shorted to the cathode, and it functions to repel secondary electrons. The pentode diagram is shown below.



The design of tetrode and pentode circuits is essentially the same as in triode circuits, with the extra complication of the additional freedom of choosing extra grid voltages, and of more complicated mathematical models for the tube characteristics.

An alternative solution to the suppressor grid comes in the **beam power tube**. This is generally; a pentode, with grid #3 replaced by a **beam-forming electrode**. (Confusingly, data sheets can refer to the same beam power tubes as tetrodes *or* pentodes.) The schematic symbol appears below.



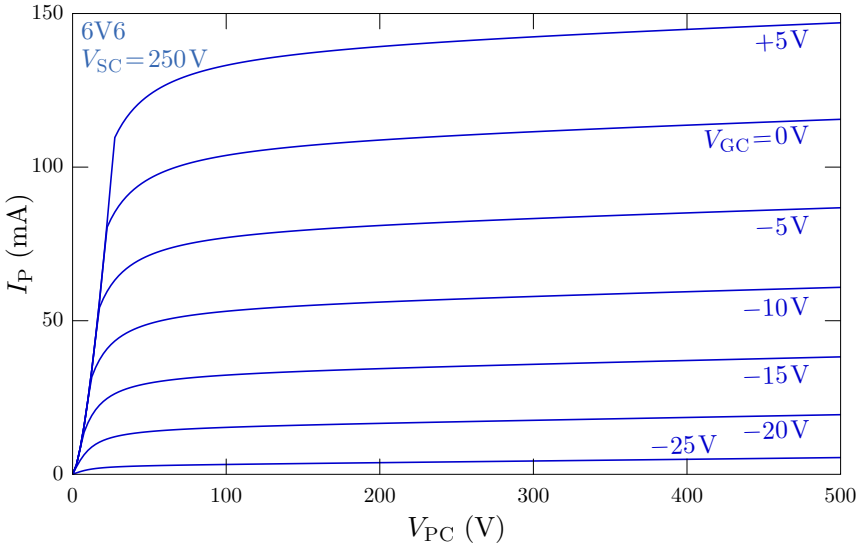
Like the suppressor grid, the beam-forming electrode is typically shorted to the cathode, and serves to direct the cathode electrons in a beam onto a particular part of the plate. Because of the controlled electron impact on the plate, such tubes can handle high currents. It also serves to suppress secondary electrons in the same way as the suppressor grid. Famous examples of beam power tubes are the 6V6 at lower powers and the 6L6 at somewhat higher powers ("6" = 6 elements: cathode, heater, plate, control grid, screen grid, beam-forming electrode). For comparison to the 12AX7 preamplifier tube, the 6V6 in typical operattion has a

plate voltage ranging from $V_{PC} = 180$–$315\,V$, has a screen voltage from $V_{G_2C} = 180$–$225\,V$, has a control-grid voltage from $V_{G_1C} = -8.5$–$13\,V$, has a transconductance of $g_m \sim 3700\,\mu\mho$, an output of 2–5.5 W, and a plate current $I_{PC} = 29$–$34\,mA$.

A naive model of the behavior of pentodes and beam power tubes would simply extend the Child–Langmuir Law in the spirit of Eq. (6.2) for triodes, to include the effect of the screen bias voltage:

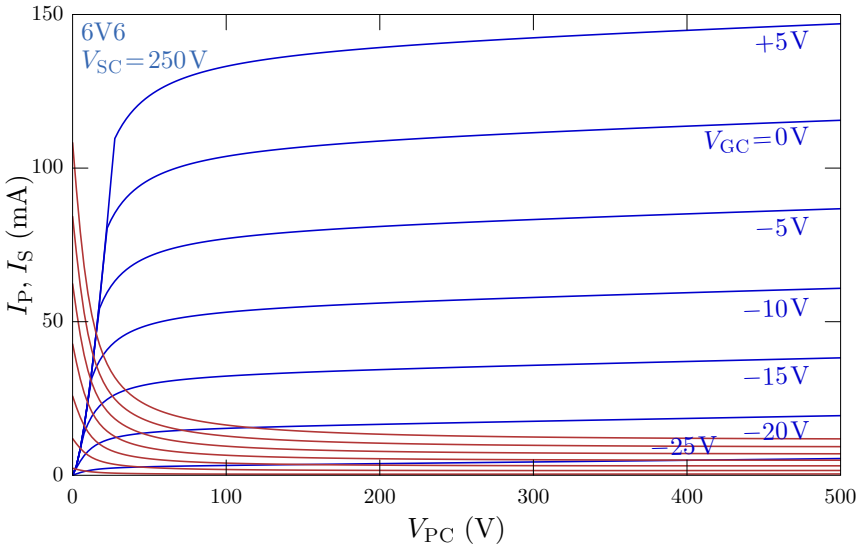
$$I_{PC} \propto (\mu_S V_{SC} + \mu_G V_{GC} + V_{PC})^{3/2} \qquad (\mu_S V_{SC} + \mu_G V_{GC} + V_{PC} > 0). \tag{6.43}$$

However, this law has a much smaller range of applicability than the analogous triode law: it does not describe the regime where I_{PC} becomes insensitive to V_{PC}, as we mentioned above for tetrodes. To visualize this, we can plot characteristic curves for the typical operation of the 6V6 beam tetrode, with a fixed screen–cathode voltage $V_{SC} = 250\,V$. A reasonable mathematical model for this is essentially Eq. (6.43), but with extra factors to enforce the "knee" behavior at larger V_{PC}.[6] Compare this to the 12AX7 curves in the figure on 132: the plate current here rises rapidly for small V_{PC} but then levels off for larger plate voltage.



Note that this diagram may be somewhat deceptive, because at small V_{PC}, the screen voltage is much higher than the plate, and so the screen can attract electrons and develop a substantial current. The figure below superimposes the plate-current curves I_S corresponding to the same gate voltages V_{GC} (ordered with increasing current in the same way as the plate-current curves).

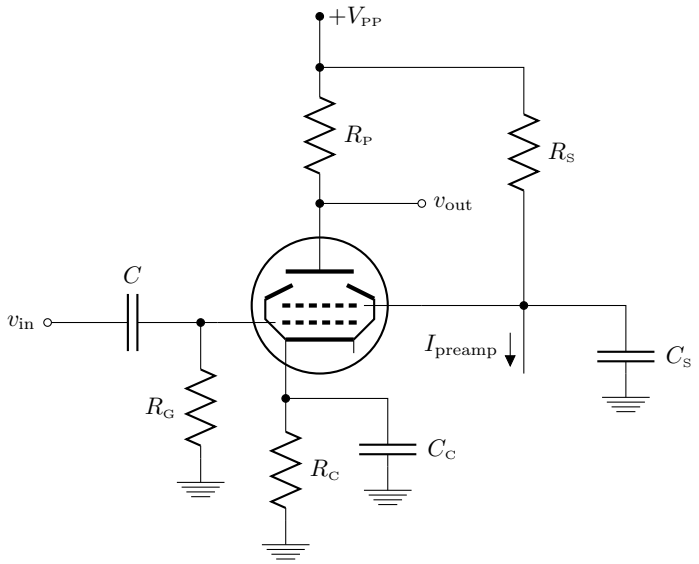
[6] The model used here comes from D. Munro, "PSpice Model: 6V6GT," `http://www.duncanamps.com/pdf/6v6spicemod.pdf`, a PSpice-language mathematical model of the tube behavior. Another model using the same equations but different parameters was posted by Robert McLean at `http://www.audiobanter.com/showthread.php?t=70465`, and this model gives similar results.



The screen current can be substantial in such a tube, but the screen is a relatively delicate device compared to the plate. For this reason, these tube circuits often employ a resistor to limit the screen current to protect and prolong the life of the screen.

6.4.1 Design Example: 6V6 Power Output Stage

As an example of a pentode-type amplifier circuit, consider the beam-power-tube amplifier power amplifier shown below, based on the 6V6 tube.



The circuit is similar to the 12AX7 preamplifier before, but now there is a screen resistor R_S that limits the screen current and voltage. A small current I_{preamp} is also tapped at the screen to power a preamplifier circuit and further drop the screen voltage.

Unfortunately, the analysis of this power amplifier is more complicated than the triode case because of the extra degree of freedom of the screen voltage. Also the screen current I_S can be significant. Since there are now three currents (plate, screen, cathode) in the tube that add as

$$I_C = I_P + I_S, \tag{6.44}$$

the load-line graphical analysis is more complicated because the voltage drops across R_P and R_C arise due to different currents:

$$V_\mathrm{PP} - V_\mathrm{PC} = I_\mathrm{P} R_\mathrm{P} + I_\mathrm{C} R_\mathrm{C}. \tag{6.45}$$

This leads to a load "line" which isn't a line anymore, because I_S is not simply related to I_P:

$$V_\mathrm{PP} - I_\mathrm{P} R_\mathrm{P} - I_\mathrm{C} R_\mathrm{C} = V_\mathrm{PC}. \tag{6.46}$$

Further, I_S determines the ground-referenced screen voltage via

$$V_\mathrm{S} = V_\mathrm{PP} - (I_\mathrm{S} + I_\mathrm{preamp}) R_\mathrm{S}, \tag{6.47}$$

but the screen–cathode voltage is also determined by the cathode bias:

$$V_\mathrm{SC} = V_\mathrm{S} - I_\mathrm{C} R_\mathrm{C}. \tag{6.48}$$

We still have the plate and control-grid bias voltages, as in the triode case:

$$\begin{aligned} V_\mathrm{PC} &= V_\mathrm{P} - I_\mathrm{C} R_\mathrm{C} \\ V_\mathrm{GC} &= -I_\mathrm{C} R_\mathrm{C}. \end{aligned} \tag{6.49}$$

This leads to a relatively complicated coupled system of equations. The graphical analysis in particular is complicated by the need to consult graphs of both the plate and cathode currents to even set the proper bias. In practice, crude approximations such as assuming I_S to be a fixed fraction of I_C are sufficient for approximate design work. But if a mathematical model of the tube is available, it is more convenient nowadays to let a compute grind out the numerical solution of these equations.

As an example, suppose we assume the following parameter set:

- $R_\mathrm{P} = 284.2\,\Omega$ (i.e., the dc resistance of the primary coil of an output transformer),
- $R_\mathrm{C} = 470\,\Omega$,
- $R_\mathrm{S} = 1\,\mathrm{k}\Omega$,
- $R_\mathrm{G} = 220\,\mathrm{k}\Omega$ (again, the grid resistor only really matters for the input impedance and to tie the grid to ground; the grid current is negligible provided it is biased negative with respect to the cathode),
- $V_\mathrm{PP} = 360\,\mathrm{V}$, and
- $I_\mathrm{preamp} = 2\,\mathrm{mA}$ to power a small preamplifier.

The numerical solution to the above equations, including the 6V6 model, gives

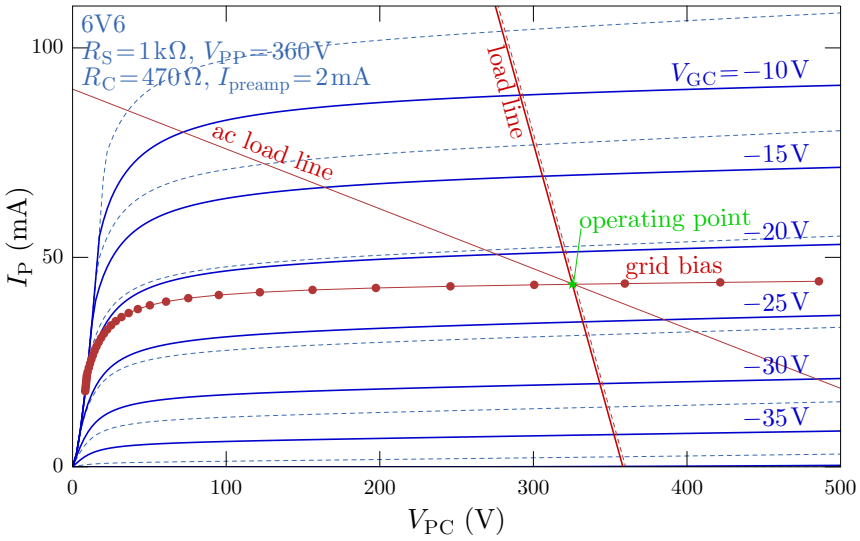
$$\begin{aligned} I_\mathrm{P} &= 43.6\,\mathrm{mA} \\ I_\mathrm{S} &= 3.75\,\mathrm{mA} \\ I_\mathrm{C} &= 47.3\,\mathrm{mA} \\ V_\mathrm{PC} &= 325\,\mathrm{V} \\ V_\mathrm{GC} &= -22.2\,\mathrm{V} \\ V_\mathrm{S} &= 354\,\mathrm{V} \\ V_\mathrm{P} &= 347\,\mathrm{V}. \end{aligned} \tag{6.50}$$

To check the tube operation, these results imply power dissipations of

$$\begin{aligned} P_\mathrm{P} &= V_\mathrm{PC} I_\mathrm{P} = 14.2\,\mathrm{W} \\ P_\mathrm{P} &= V_\mathrm{SC} I_\mathrm{S} = 1.25\,\mathrm{W} \end{aligned} \tag{6.51}$$

at the plate and screen, respectively. The 6V6 has specified maximum powers of 12 W and 2 W, respectively, so this design is pushing the limits of the plate power. (If the plate power exceeds the specification by too much, the tube will "red plate," or the anode will glow red with heat, dramatically shortening the tube's life.)

All this and more is illustrated in the plot below. The characteristic curves for this arrangement are shown as solid curves; the dashed curves show the corresponding characteristics when V_{SC} is held at a constant 330 V, illustrating the importance of numerically solving for the correct V_{SC} in the analysis.



The (dc) load curve is plotted from Eq. (6.46); the dashed counterpart is the load line obtained by ignoring the small contribution of I_S. Clearly the screen current makes a difference, but here it doesn't change much to ignore it.

The grid-bias curve is more difficult to construct graphically here because it only intersects one characteristic curve (for $V_{GC} = -20\,V$). The plotted points are separated by $\Delta V_{GC} = 0.1\,V$, from $V_{GC} = -19\,V$ to $-22.5\,V$, from left to right. Recall that this curve is determined by the cathode current, which is determined by the plate and screen voltages. The intersection with the load line again gives the operation point.

A different, ac load line is also shown. Since R_P corresponds to the primary winding of an output transformer, it has a different ac response than the dc resistance. The plotted curve assumes a nominal ac impedance of 7 kΩ. This load line shows how changes to the gate voltage translate into output-current changes. Note that if V_{GC} rises much above $-10\,V$, the curves bunch together, resulting in clipping of the signal; the same is true if V_{GC} goes below $-35\,V$ or so.

Computing the ac gain directly is somewhat easier than setting the dc bias, because the bypass capacitors allow us to assume that the cathode voltage V_C and screen voltages V_S and V_{SC} are fixed. The transconductance can be evaluated by numerically differentiating the tube model. The result here is

$$g_m = \frac{\partial I_P}{\partial V_{GC}} = 4330\,\mu\mho, \tag{6.52}$$

which is similar to values specified in the data sheet. The plate resistance then follows from adapting the definition (6.15) to read

$$\frac{1}{r_p} := \frac{\partial I_P}{\partial V_{PC}}. \tag{6.53}$$

This derivative can be evaluated numerically as well, and comes to

$$r_p = 72.5\,k\Omega. \tag{6.54}$$

Because of the extra dependence on the screen voltage, the relation $g_m r_p = \mu$ of Eq. (6.20) no longer holds exactly. For example, the 6V6 model assumes a grid μ of 390 (and a screen μ of 43), but $g_m r_p = 314$ from the numerical calculations

The ac voltage gain from Eq. (6.26) is then

$$G = -\frac{g_m r_p Z_P}{r_p + Z_P} = -27.7, \tag{6.55}$$

after restoring $g_m r_p$ in place of μ and taking R_P to be the ac impedance Z_P. Note that the "naïve" result $G = -g_m Z_P$ from Eq. (6.14) works reasonably well here, predicting $G = -30.3$, because the plate resistance is indeed substantially larger than the plate impedance, by more than a factor of 20. In any case, the gain here is lower than in the triode preamplifier example.

6.4.2 Triode Connection

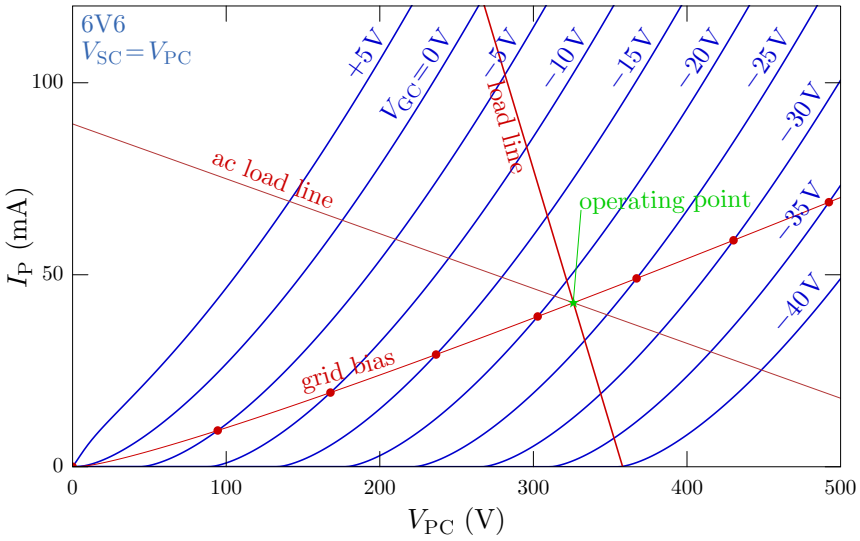
The characteristic curves in the triode-amplifier example of Section 6.2.3 appeared quite different from the beam-power-tetrode example of Section 6.4.1. Roughly speaking the triode is operated in a regime where it behaves like a variable resistor, controlled by the grid voltage. This is evident in the gain expressions (6.25) and (6.26), where the plate resistance (along with the magnification) determine the tube's contribution to the amplifier gain. The pentode-type circuits, however, when the screen is held at a fixed potential, are different: they typically run in a regime with large plate resistance, making the tube output behave more like a current source (since the plate current is insensitive to V_{PC}). In this regime the tube's contribution to the amplifier gain comes in the form of the transconductance, as in Eq. (6.28).

Pentodes (and beam power tubes) work well enough, but sometimes triode characteristics are more desirable, for example for their distortion or low-noise properties. One common trick for achieving this with a pentode-type tube is the **triode connection**, where the screen is shorted to the plate (typically this is done via a resistor of $\sim 100\,\Omega$, which limits screen current and suppresses high-frequency instabilities).

As an example, consider the same power-amplifier stage as in the previous section, but with a triode-connected 6V6 (with a direct plate–screen short for simplicity). The analysis is essentially the same, except Eq. (6.47) is replaced by the condition $V_S = V_P$. The dc operating point turns out to be nearly the same (all the values below are within a few percent of the previous values):

$$\begin{aligned} I_P &= 42.7\,\text{mA} \\ I_S &= 3.68\,\text{mA} \\ I_C &= 46.4\,\text{mA} \\ V_{PC} &= 326\,\text{V} \\ V_{GC} &= -21.8\,\text{V} \\ V_S &= 348\,\text{V} \\ V_P &= 348\,\text{V}. \end{aligned} \tag{6.56}$$

The characteristic curves are shown below, with the same load curves, and the appropriate grid-bias curve.



As might be expected, the curves here resemble more those of the 12AX7 (in the plot on p. 132) than the pentode-type curves (from the plot on p. 138).

For the ac analysis, the transconductance is similar to the pentode-operation case, with

$$g_m = 4310\,\mu\mho \tag{6.57}$$

(formerly $g_m = 4330$), but the plate resistance drops dramatically to

$$r_p = 2.1\,\text{k}\Omega \tag{6.58}$$

(from a former 72.5 kΩ), because it measures the plate-current change caused by variation in the plate voltage, which also varies the screen voltage (the screen voltage has a stronger effect via its magnification). The resulting gain is then

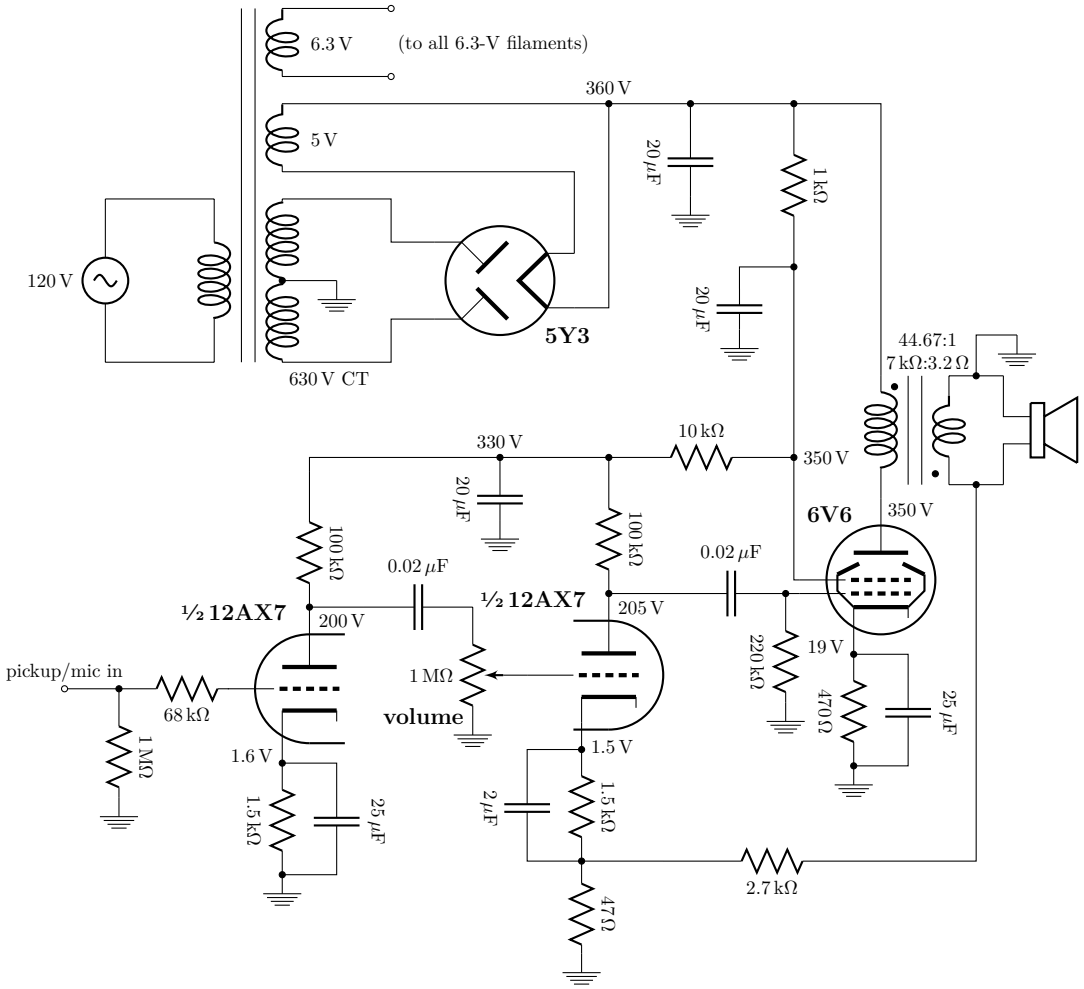
$$G = -\frac{g_m r_p Z_P}{r_p + Z_P} = -6.8, \tag{6.59}$$

(reduced from −27.7). This is fairly typical behavior obtained from triode-connecting a pentode-type tube: much-decreased gain, but with (potentially) more desirable characteristics. For example, along the ac load line, the constant-V_{GC} curves bunch together at the extremes, indicating distortion. In the triode case, they appear to remain evenly spread out, which is one consequence of the lower gain. However, another characteristic of the triode configuration is the distortion tends to turn on smoothly, which leads to a "warmer" triode sound, compared to a "clean" and possibly "harsh" pentode sound in audio circuits.

6.5 Circuit Practice: Simple Tube Amplifier

As an example of a more complete vacuum-tube circuit, consider below an amplifier based on three tubes for electric guitar or microphone. This amplifier design is similar to various "Champ" amplifiers made by Fender Musical Instruments Corp., and the voltages are nominal values (with 20% stated acceptable tolerance) reported on schematics of those instruments.[7]

[7] Gerald Weber, *A Desktop Reference of Hip Vintage Guitar Amps* (Kendrick Books, 1996) (ISBN: 0964106000).



6.5.1 First Preamplifier Stage

Starting with the input at the first triode amplifier, we can see that this is essentially the same triode amplifier as in the diagram on p. 128, but with a slightly different input network. Since the intended source is a guitar pickup or dynamic microphone, which works by induction in a sensing coil, no ac-coupling capacitor is needed. However, the input is tied to ground by a 1-MΩ impedance in case no source is connected, and any input current is limited by the 68-kΩ resistor in case some input connection tries to bring the grid excessively positive. Because the cathode-bias resistor is bypassed with a capacitor, above about 4 Hz the gain expression (6.26) applies, and the numerical estimate (6.42) of $G = -65$ from the triode example applies here. Note that plate and cathode resistors, as well as V_{PP} in the triode example were chosen to match this design. Actually, though, the 1-MΩ resistor in the second stage gives an ac load for the amplifier, so for the gain calculation we should use $R_P = 90.9\,\text{k}\Omega$, leading to a better estimate $G = -63$. The grid-bias voltage of -1.6 V from Eqs. (6.38) is in good agreement with the nominal value for the cathode voltage. The plate voltage of $V_P = V_{PC} - V_{GC} = 221\,\text{V} + 1.6\,\text{V} = 223\,\text{V}$ is a bit over 10% over the nominal value quoted of 200 V. Thus the schematic implies a slightly higher plate current than the model (1.3 mA vs. 1 mA).

Assuming the nominal values from the schematic, it is also possible to estimate the gain from the RCA data sheet.[8] Using $V_{PC} = 200\,\text{V}$ and $V_{GC} = -1.6\,\text{V}$ yields $r_p = 62\,\text{k}\Omega$, $g_m = 1600\,\mu\mho$, and $\mu = 100$. From these values, Eq. (6.26) gives a gain of $G = -65$ (including the load), which is reasonably close (within 5%) of the above estimate.

[8] Available at `http://drtube.com/datasheets/12ax7-rca1962.pdf`, see the plot on the last page.

6.5.2 Second Preamplifier Stage

The second-stage preamplifier, after the volume control, is essentially identical to the first, except for the addition of the 47-Ω resistor and the feedback path from the output transformer. We will save the feedback for later, but the extra resistance yields modified values of $V_{PC} = 226\,\text{V}$ (vs. 225 V) and $V_{GC} = -1.59\,\text{V}$ (vs. $-1.61\,\text{V}$). Again, these are close to the nominal values. The gain expression (6.25) gives $G = -54$, using $R_P = 68.8\,\text{k}\Omega$ (including the 220-kΩ input resistor of the power stage) and $R_C = 47\,\Omega$. The data sheet for these conditions gives $r_p = 60\,\text{k}\Omega$, $g_m = 1650\,\mu\text{℧}$, and $\mu = 101$, giving a gain of $G = -52$.

6.5.3 Power Amplifier Stage

The power-amplifier stage is intended to match the prototype power-amplifier circuit of Section section:6V6amp, including the 1 mA of current that supplies each preamplifier stage. The output transformer acts as a load; the transformer is a Hammond 1750C.[9] The primary dc resistance is specified as $284.2\,\Omega$, but can vary considerably (one example measured $267\,\Omega$). The transformer specifies an ac impedance ratio of 7000 : 3.2, which means that a 3.2-Ω speaker load appears as a 7-kΩ load at the 6V6 plate. (The ac impedances are nominal values, varying widely with frequency because of the frequency response of the transformer and especially of the speaker.) The ac gain for this stage is $G = -27.7$.

6.5.4 Feedback Loop

The main remaining feature of the circuit, besides the power supply, is the 2.7-kΩ resistor connecting the transformer secondary to the cathode resistor chain of the second premplifier stage. This provides overall *negative* feedback in the circuit. That the feedback is negative is not entirely obvious, and so it's worth quickly tracing it through. When the input voltage on the second preamp increases; the input to power tube goes down (owing to the negative gain of the common-cathode amp); the plate voltage on power tube goes up; this reduces current in output-transformer primary (i.e., the change in current points towards the dot); this increases current in secondary (i.e., the change in current again points to dot); and finally this increases cathode voltage at the preamp, which counteracts effect of input (in a way similar to an unbypassed cathode resistor).

To work out the feedback gain, note that the transformer *voltage* ratio is $\eta_T = \sqrt{3.2/7000}$. (Recall that the voltage ratio of a transformer is the same as the turns ratio, while the current ratio is the inverse of the turns ratio; hence the impedance $Z = V/I$ ratio is the *square* of the turns ratio.) Then the net open-loop gain from the input to the second preamplifier to speaker is

$$G_{net} = \eta_T G_{pre,2} G_{power} = \sqrt{3.2/7000} \times (-54) \times (-27.7) = 32, \tag{6.60}$$

where $G_{pre,2}$ is the voltage gain of the second preamplifier, and G_{power} is the voltage gain of the power-amplifier stage. To account for the effect of the feedback, we can replace the 47-Ω and 2.7-kΩ resistors by the $R_f = 46\,\Omega$ parallel (Thévenin) resistance and a Thévenin voltage of $v_f = \eta_{div} G_{net} v_{in}$, where v_{in} is the input to the second preamplifier stage, and $\eta_{div} = 0.0171$ is the voltage-divider ratio due to the 47-Ω and 2.7-kΩ resistors.

With this modification, Eq. (6.23) is modified to read (setting $R_C = 0$ since it is bypassed with a capacitor

$$\begin{aligned} i_{PC} &= g_m(v_{in} - v_C) - \frac{R_P + R_f}{r_p} i_{PC} \\ &= g_m(v_{in} - i_{PC}R_f - v_f) - \frac{R_P + R_f}{r_p} i_{PC} \\ &= g_m(v_{in} - i_{PC}R_f - \eta_{div}G_{net}v_{in}) - \frac{R_P + R_f}{r_p} i_{PC}. \end{aligned} \tag{6.61}$$

Solving for i_{PC} gives

$$i_{PC} = \frac{g_m(1 - \eta_{div}G_{net})\, v_{in}}{1 + g_m R_f + (R_P + R_f)/r_p}. \tag{6.62}$$

[9] The transformer data sheet: `http://www.hammondmfg.com/pdf/EDB1750C.pdf`.

Then with $v_{\text{out}} = -i_{\text{PC}} R_{\text{P}}$ the closed-loop ac voltage gain for the preamplifier becomes

$$G_{\text{pre,2,CL}} = -\frac{\mu(1 - \eta_{\text{div}} G_{\text{net}}) R_{\text{P}}}{r_{\text{p}} + (\mu + 1) R_{\text{f}} + R_{\text{P}}} = -24, \tag{6.63}$$

again using the parameters $g_{\text{m}} = 1880\,\mu\mho$, $\eta_{\text{div}} = 0.0171$, $G_{\text{net}} = 32$, $R_{\text{P}} = 68.8\,\text{k}\Omega$, $R_{\text{C}} = 1.5\,\text{k}\Omega$, $r_{\text{p}} = 5.19\,\text{k}\Omega$, $\mu = 100$, and $R_{\text{f}} = 46\,\Omega$. The preamplifier-stage gain here is reduced from the open-loop value of $G_{\text{pre,2}} = -54$. The reduction in gain is of course accompanied by a slight improvement in gain flatness with frequency and reduction in distortion (see Section 7.7). However, the improvements are not extreme, since the change in gain is only a factor of 2.

With the negative feedback, the gain from the second preamplifier input to the speaker after the volume control is then $G_{\text{net,CL}} = \eta_{\text{T}} G_{\text{pre,2,CL}} G_{\text{power}} = 14$. The gain including the first preamplifier stage ($G_{\text{pre,1}} = -63$), when the volume control is maximized, is $G = 900$ (dropping the sign, which doesn't matter anymore). The gain from first input to the transformer primary (i.e., the 6V6 plate voltage) is 4.2×10^4.

To give a sense of scale of this voltage, a 10-mV input signal gives a much larger signal at the input of the power tube of $(10\,\text{mV}) \times 24 \times 63 = 15\,\text{V}$. This is comparable to the gate bias voltage of the 6V6. Thus, at the maximum gain setting, the amplifier is already at the threshold of overdriving the input of the power tube, leading to distortion. The distortion can be desirable for electric-guitar amplifiers, and is easily attainable here: maagnetic pickups for electric guitars have typical maximum signal levels in the range of 100–500 mV. Dynamic microphones tend to have smaller, mV-level signals, which would lead to much less distortion in this circuit.

6.5.5 Power Supply

The power supply is a straightforward variation of the full-wave rectifier on p. 125. However, the voltage is not entirely obvious. The plate winding of the power transformer[10] is specified at 630 V, center-tapped, for 100 mA load current. This results in 315 V to either anode of the 5Y3, but this is rms: the peak anode voltage is 445 V. The rectifier tube drops 50 V at 125 mA dc forward current, but this does not explain the 360 V nominal output voltage in the diagram. The transformer data sheet, however, conveniently specifies a 363 V (dc) output using a 5Y3 tube at 100 mA, implying a larger voltage drop than suggested by the GE data sheet.[11] The answer to this is on page 4 of the 5Y3 data sheet, which shows the loaded, rectified output voltage, which has more "sag" when filtered by a capacitor: the plot on the data sheet shows about 340 V dc output for a 100-mA dc output current and 315-V rms input (with a 20-μF filter capacitor), and about 380 V output for a 50-mA dc output. These numbers are reasonably consistent with the nominal voltage in the schematic, and reflect differences in characteristics of tubes available from different manufacturers.

The rectifier output is then filtered by a 20-μF capacitor, as in the 5Y3 data sheet, and this supplies the plate voltage for the 6V6. A 1-kΩ resistor feeds the screen from the same voltage, as we analyzed in the 6V6 power-amplifier example. This drops the screen voltage down to the plate-voltage level. The screen voltage is also bypassed by another 20-μF capacitor, decoupling the screen and plate. The screen voltage then supplies the 2 mA to the 12AX7 plates, with an appropriate voltage drop across a 10-kΩ resistor. The preamplifier plate voltages are bypassed by yet another another 20-μF capacitor, decoupling the preamplifier tubes from the power tube and from each other.

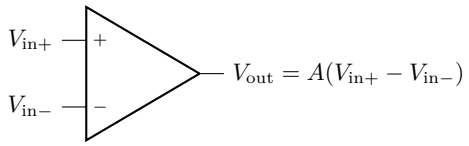
[10] Magnetic Components 40-18027, `http://www.classictone.net/40-18027.pdf`.
[11] `http://www.r-type.org/pdfs/5y3gt.pdf`

Chapter 7

Operational Amplifiers

7.1 Op-Amp Basics

We have already talked a bit about differential amplifiers, when we analyzed the transistor differential amplifier. The differential amplifier is an important building block in its own right, and we will spend a fair amount of time looking at them. In particular, the **operational amplifier (op-amp)** is a handy, handy device in analog-circuit design. Recall that the differential amplifier ideally subtracts the input voltages, and multiplies the difference by a gain factor (here, A) to produce an output signal.



The operational amplifier is basically a differential amplifier, but with large gain (with larger=better, where op-amps are concerned). The gain A here is called the **open-loop gain** of the op-amp, for reasons that will become more apparent soon. For real op-amps, gains typically range from around 46 dB on the low end (for low-quality amplifiers, or amplifiers where other engineering considerations compromised the gain), to around 140 dB on the high end. (Remember that 20 dB corresponds to a factor of 10 in voltage, so 140 dB means $A = 10^7$.) For more specific examples, let's summarize the gains of a few classic op-amps:

- **741C**: has BJT inputs, $A = 86$ dB, a famous, old op-amp that is cheap, and not so great anymore (there are better choices, even among cheap op-amps).
- **LF411**: has JFET inputs, $A = 88$ dB, a cheap op-amp, which is really not bad, and a good "default op-amp" in noncritical applications (Horowitz and Hill call this their "jelly bean").
- **OPA111B**: has JFET inputs, $A = 120$ dB, a venerable, precision, low-noise op-amp, but *expensive*.

7.1.1 Usage: Open-Loop

There are two basic ways to use an op-amp: in the first, we take advantage of the high gain, and in the second, we throw away some of the high gain. The first application, where we use the full gain of the op-amp, is called **open-loop mode**, and the op-amp behaves as a **comparator**. To explain this, note that the op-amp is an active device, and requires a power supply; often op-amps are powered by split ±15-V power supplies, so the output can go either positive or negative. The output, of course, cannot exceed the power-supply voltages (and can typically the output range, or **output swing**, of the op-amp is a volt or two *less* than the supply range). Now when the output formula,

$$V_{\text{out}} = A(V_{\text{in}+} - V_{\text{in}-}) \tag{7.1}$$

predicts that V_{out} should be outside the supply-voltage range, what *really* happens is that the output **rails**; for example, if the output *should* be +50 V, but the power supplies are ±15 V, the output will *rail* at +15 V (or more likely a bit lower, say around 14.3 V.

Then the **comparator** action of an op-amp is as follows:

1. If $V_{\text{in}+} > V_{\text{in}-}$ by at least a few mV, then the V_{out} rails at the *positive* supply voltage.
2. If $V_{\text{in}+} < V_{\text{in}-}$ by at least a few mV, then the V_{out} rails at the *negative* supply voltage.

That is, the op-amp *compares* the input voltages, and swings the output to the appropriate rail to indicate which one is bigger. There are specialized op-amps, called **comparators**, that are optimized to do this, and we will consider comparator circuits in more detail later. For now, note that regular op-amps can be used in this way, though this is not the most common usage.

7.1.2 Usage: Closed-Loop

In **closed-loop mode** or **negative-feedback mode**, op-amps have some connection from the output to the inverting (−) input. The net effect is to reduce the gain. Why bother to have an op-amp with extremely high open-loop gain only to reduce it in closed-loop mode? Well, it turns out that when you do this, the resulting circuit behaves *well*, in the sense that its behavior will be (mostly) independent of the device properties. This becomes more true with increasing open-loop gain. This trick of using **negative feedback** is a really nice trick, and opens up a lot of possibilities for cool circuits.

7.2 Op-Amp "Golden Rules"

In the simplest method for analyzing op-amp circuits, we will assume that we are dealing with an *ideal* op-amp. This means that we will assume the following two rules:

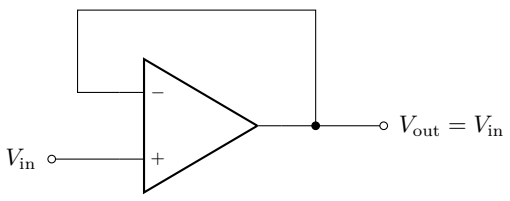
1. No current flows into or out of the inputs. (Current *can* of course flow into or out of the *output*, as well as the power-supply terminals, which we haven't bothered to label thus far.)
2. Either $V_{\text{in}+} = V_{\text{in}-}$, or the output is railed. Basically we are assuming that the open-loop gain A is so large, that the only way for the output to *not* be railed is for $V_{\text{in}+}$ and $V_{\text{in}-}$ to be basically the same, because
$$V_{\text{in}+} = V_{\text{in}-} = \frac{V_{\text{out}}}{A} \approx 0. \tag{7.2}$$
Remember that for the output not to rail, there must generally be negative feedback. In this case *the output does whatever it needs to do* to make sure the input voltages are approximately equal.

Using just these two rules, we can analyze the basic behavior of most op-amp circuits.

7.3 Basic Op-Amp Circuits

7.3.1 Unity-Gain Buffer/Follower

The first circuit is simple: the **unity-gain buffer**, or **voltage follower**. Here, $V_{\text{out}} = V_{\text{in}}$, and this is mostly useful to buffer a high output impedance or a low input impedance. The improvement here on the transistor followers is that this circuit works for dc voltages, not just biased ac voltages.



To analyze this, note that the output is shorted to the inverting input, providing negative feedback. Then

$$V_{\text{in}-} = V_{\text{out}}. \tag{7.3}$$

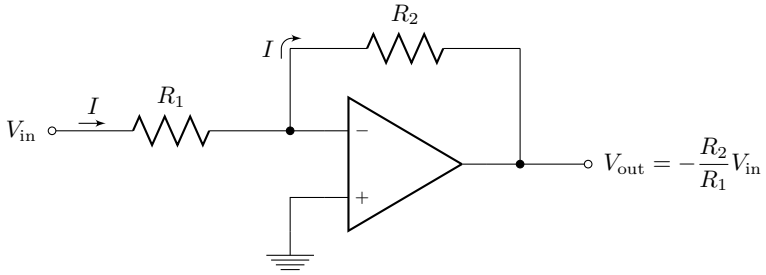
But the second op-amp rule says that $V_{\text{in}+} = V_{\text{in}-}$, so $V_{\text{in}+} = V_{\text{in}}$, or

$$V_{\text{out}} = V_{\text{in}-}. \tag{7.4}$$
(unity-gain buffer)

Hence, this circuit has unity gain for dc signals. We will return later to the question of the input and output impedances of this circuit, but we expect the input impedance to be high, considering that ideally no current flows into the input. But typically, we expect input impedances of $\sim 10^8\,\Omega$ for op-amps with BJT inputs, and $\sim 10^{12}\,\Omega$ for op-amps with JFET inputs. The output impedances can be in the range of $m\Omega$ for good op-amps. So this circuit is handy for buffering high-impedance sensors, for example. Another common use is to derive reference voltages in circuits: For example, if you need a voltage of +5-V somewhere in a circuit, you can use a voltage divider between ±15-V power supplies, and buffer the output of the divider so you don't have to worry about loading it down.

7.3.2 Inverting Amplifier

The **inverting amplifier** is useful as a basic amplifier *with* gain, and not only because it has many useful variations.



To analyze this circuit, assume a current I flows into the input. Golden rule 2 says that the inverting-input voltage $V_{\text{in}-}$ must be zero, so

$$I = \frac{V_{\text{in}}}{R_1}. \tag{7.5}$$

Then according to the first rule, no current goes into the input, so it must all go through the feedback resistor R_2. Then applying Ohm's law across the feedback resistor,

$$0\,\text{V} - V_{\text{out}} = IR_2, \tag{7.6}$$

or putting in the previous expression for I and solving for the output voltage,

$$V_{\text{out}} = -\frac{R_2}{R_1}V_{\text{in}}. \tag{7.7}$$
(inverting amplifier)

Defining the **closed-loop gain** G as

$$G := \frac{V_{\text{out}}}{V_{\text{in}}}, \tag{7.8}$$
(inverting amplifier)

we can write the closed-loop gain of the inverting amplifier as

$$G = -\frac{R_2}{R_1}. \tag{7.9}$$
(closed-loop gain, inverting amplifier)

This must be smaller than the open-loop gain, otherwise the assumption of negative feedback breaks down.

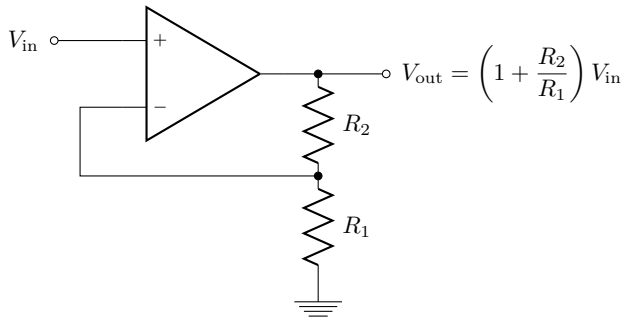
Now it's useful to try to get an idea of how feedback works in this circuit to maintain the advertised output voltage. Let's take the simple case $R_1 = R_2$, for unity gain ($G = -1$). Suppose the output voltage is initially zero, and we suddenly introduce an input $V_{\text{in}} = 1\,\text{V}$.

- Then the R_1–R_2 pair acts as a voltage divider, putting the voltage $V_{\text{in}-}$ at the inverting input at 0.5 V. Since the inverting input is above the (grounded) noninverting input, the output "wants" to go *down* in voltage.
- Suppose the output *overshoots* to, say, $-2\,\text{V}$. Then $V_{\text{in}-} = -0.5\,\text{V}$, and the inverting input is *below* the noninverting input, so the output wants to go *up*.
- The only output that causes the inputs to be balanced is the "proper" output of $-1\,\text{V}$. Any deviation away from this makes the op-amp want to drive the output towards this value. So not only does it satisfy the golden rules, but it is **stable**, in that the op-amp will correct for deviations away from this value.

This one of the powers of negative feedback: robustness to deviations from the proper output (e.g., due to power-supply noise). If the feedback is not negative, then more complicated behaviors like oscillation may result; we will return to this useful case later.

7.3.3 Noninverting Amplifier

It is also possible to build a **noninverting amplifier** with gain, as shown below.



Using that R_1 and R_2 form a voltage divider,

$$V_{\text{in}} = V_{\text{in}+} = V_{\text{in}-} = \frac{R_1}{R_1 + R_2} V_{\text{out}}. \tag{7.10}$$

Thus, the closed-loop gain is

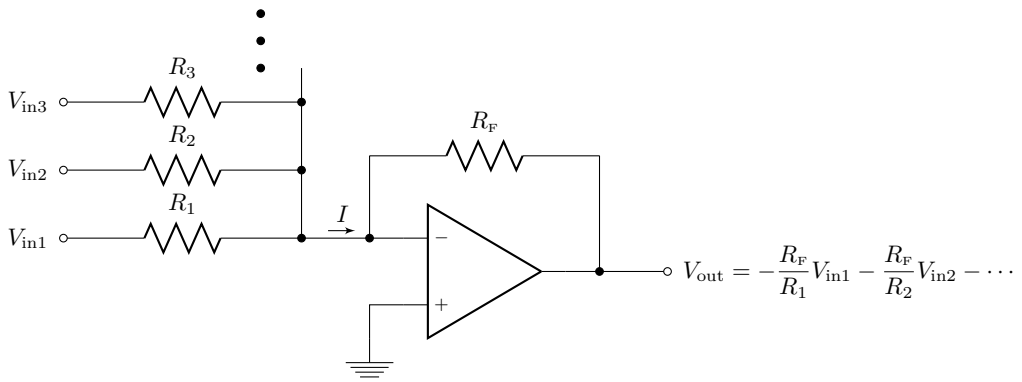
$$G = \frac{R_1 + R_2}{R_1} = 1 + \frac{R_2}{R_1}.$$

(closed-loop gain, noninverting amplifier) (7.11)

Note that in the inverting-amplifier case, sub-unity gains are possible if $R_2 < R_1$, but here, the smallest possible gain is unity. Like the unity-gain buffer, this circuit enjoys high input impedance; the input impedance in the inverting case is just R_1.

7.3.4 Summing (Inverting) Amplifier

A useful variation on the inverting amplifier is the **summing amplifier**, which combines multiple input voltages with different gains to obtain the output. (Note that you *can't* easily add voltages in passive circuits, which makes this circuit useful.)



The key to this circuit is that each input voltage and input resistor makes a current; at the big junction, all these currents add and go through the feedback resistor R_F. That is, the total current I is

$$I = \frac{V_{\text{in1}}}{R_1} + \frac{V_{\text{in2}}}{R_2} + \frac{V_{\text{in3}}}{R_3} + \cdots . \tag{7.12}$$

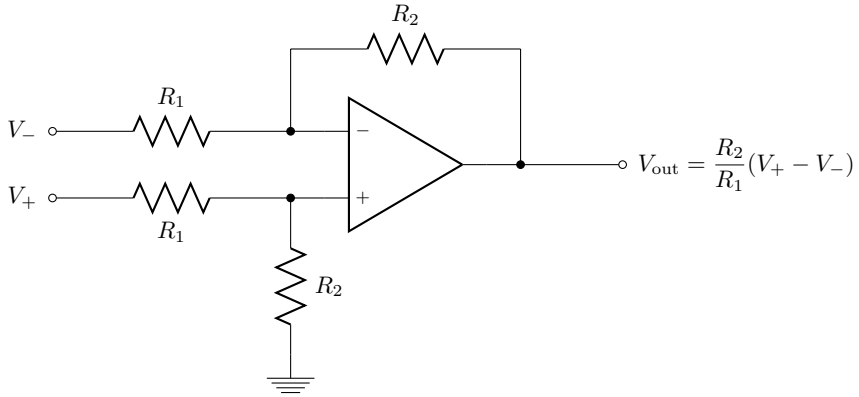
Then the output voltage is

$$V_{\text{out}} = -IR_F = -\frac{R_F}{R_1}V_{\text{in1}} - \frac{R_F}{R_2}V_{\text{in2}} - \frac{R_F}{R_3}V_{\text{in3}} - \cdots . \tag{7.13}$$
(summing amplifier)

The inverting nature of the amplifier can be inconvenient, but easily fixed by following up with another inverting amplifier.

7.3.5 Circuit Practice: Differential Amplifier

Another classic operational amplifier is shown below. This takes the difference of two input signals, and implements a closed-loop gain given by the ratio of resistors. Normally, we save the circuit practice for the end of the chapter, but here you should work this circuit out right away to review the concepts thus far before going on.



This circuit relies on the resistor values being well-matched for accuracy. Show that this circuit behaves as advertised.

Solution. First, we have a voltage divider at the noninverting input:

$$V_{\text{in}+} = \frac{R_2}{R_1 + R_2}V_+. \tag{7.14}$$

Similarly, we have a voltage divider between two voltages at the inverting input:

$$V_{\text{in}-} = \frac{R_2}{R_1 + R_2}V_- + \frac{R_1}{R_1 + R_2}V_{\text{out}}. \tag{7.15}$$

Setting these two voltages equal, we get

$$R_2 V_+ = R_2 V_- + R_1 V_{\text{out}}, \tag{7.16}$$

or

$$V_{\text{out}} = \frac{R_2}{R_1}(V_+ - V_-), \tag{7.17}$$

as desired.

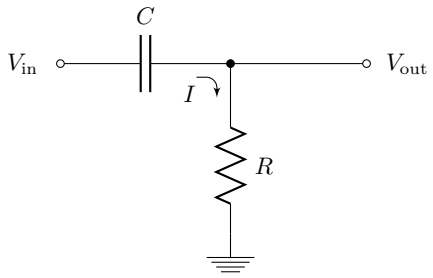
7.4 Op-Amp Filters

When we were studying *passive*, linear circuits with resistors, capacitors, and inductors, we saw that the reactive elements (capacitors and inductors) acted like frequency-dependent resistors. The resulting circuits attenuated the input signal in a frequency-dependent way, leading to passive **filters** for signals. We can do the same thing with op-amps: they open up new possibilities as well as straightforward improvements on the passive circuits.

The most important and fundamental op-amp filters are the op-amp versions of the passive integrator and differentiator. The op-amp versions have their own problems, but mainly because they have overall much more ideal behavior.

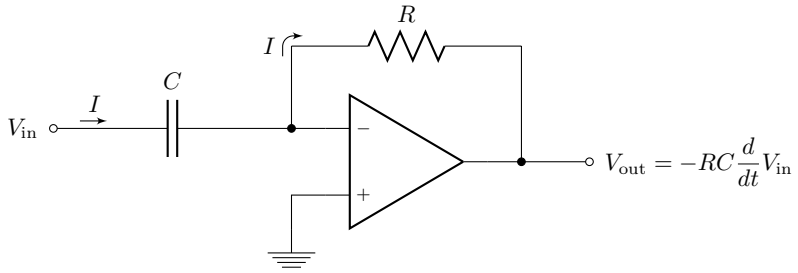
7.4.1 Op-Amp Differentiator

Recall the passive differentiator (Section 2.2.2), shown below.



Remember that this works intuitively as follows: The voltage across the capacitor is proportional to the charge ($Q = CV$). Current I flows from the capacitor through the resistor, where current is the derivative of charge ($I = dQ/dt$). The resistor converts current to voltage ($V = IR$), so the output is the derivative of the input. *But*, for the output to be the derivative of the input, we had to assume here that the output voltage is small, otherwise the voltage across the capacitor is $V_{\text{in}} - V_{\text{out}}$, not merely V_{in}.

We can improve this, essentially by using the op-amp to decouple the capacitor and resistor voltages as follows.



Intuitively, the current I flows in through the input, and the voltage across the capacitor is $V_{\text{in}} - V_{\text{in}-} = V_{\text{in}}$, enforced by the op-amp. This same current I gets converted to a voltage, via the feedback resistor.

To show this quantitatively, let's use the fact that at fixed frequency ω, this is just an inverting amplifier. The gain is $G = -R_2/R_1$, where we should replace R_2 by R, and R_1 by $X_C = i/\omega C$. Then

$$V_{\text{out}} = -\frac{R}{X_C} V_{\text{in}} = i\omega RC V_{\text{in}}. \tag{7.18}$$

Now remember that at fixed frequency ω, we can identify $d/dt \equiv -i\omega$, so

$$V_{\text{out}} = -RC\frac{d}{dt} V_{\text{in}}. \tag{7.19}$$
(op-amp differentiator)

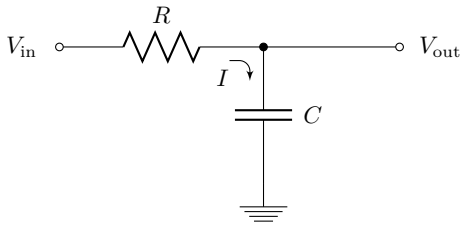
Note that the RC time and the time derivative conspire to make the voltage units come out right. Note also that the frequency-dependent gain

$$G(\omega) = i\omega RC \tag{7.20}$$
(gain, op-amp differentiator)

always increases with frequency, whereas the passive differentiator (high-pass filter) had a gain (transfer function) that leveled off at unity for frequencies above the $\omega_{3\text{dB}}$ (Section 2.3.7).

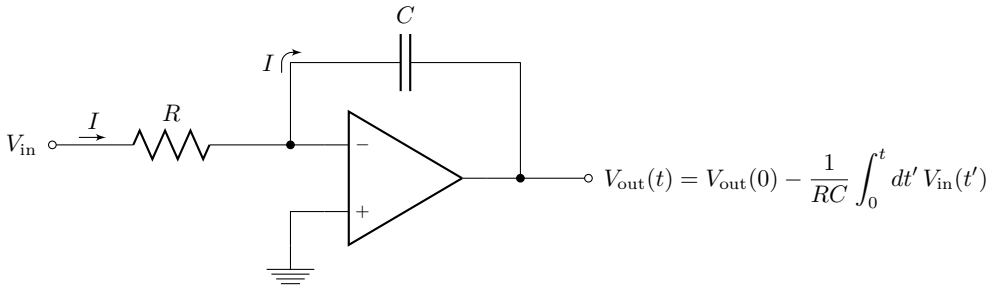
7.4.2 Op-Amp Integrator

The passive and active (op-amp) integrators are similar. Recall the passive integrator below (Section 2.2.1).



Here, the resistor converts the input voltage to a current ($V_{\text{in}} = IR$), and the capacitor develops a voltage proportional to charge ($Q = CV$), which is the integral of current. Hence the output voltage is the integral of the input voltage. But we had to assume small V_{out}, so that the voltage across the resistor is $V_{\text{in}} - V_{\text{out}} \approx V_{\text{in}}$.

Again, the op-amp helps here by decoupling the capacitor and resistor voltages, while connecting their currents, by maintaing the inverting input at virtual ground. The op-amp integrator is shown below.



This works more like we said: The input resistor converts the voltage V_{in} to current I ($V_{\text{in}} = IR$, with no need for small V_{out}), and the capacitor integrates the current to store charge, which produces an output voltage.

For the quantitative analysis, we again use the inverting-amplifier gain $G = -R_2/R_1$, with R_2 replaced by X_C, and R_1 replaced by R. Then

$$V_{\text{out}} = -\frac{X_C}{R} V_{\text{in}} = -\frac{i}{\omega RC} V_{\text{in}}, \tag{7.21}$$

so that the frequency-dependent gain is

$$G(\omega) = -\frac{i}{\omega RC}. \tag{7.22}$$
(op-amp integrator gain)

Rearranging the factor of ω,

$$-i\omega V_{\text{out}} = -\frac{1}{RC} V_{\text{in}}. \tag{7.23}$$

Then using $d/dt \equiv -i\omega$,

$$\frac{d}{dt} V_{\text{out}} = -\frac{1}{RC} V_{\text{in}}. \tag{7.24}$$

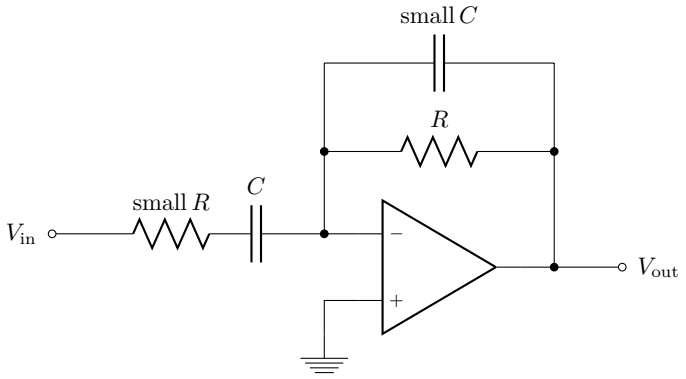
Integrating, we have

$$V_{\text{out}}(t) = V_{\text{out}}(0) - \frac{1}{RC} \int_0^t dt'\, V_{\text{in}}(t'). \tag{7.25}$$
(op-amp integrator)

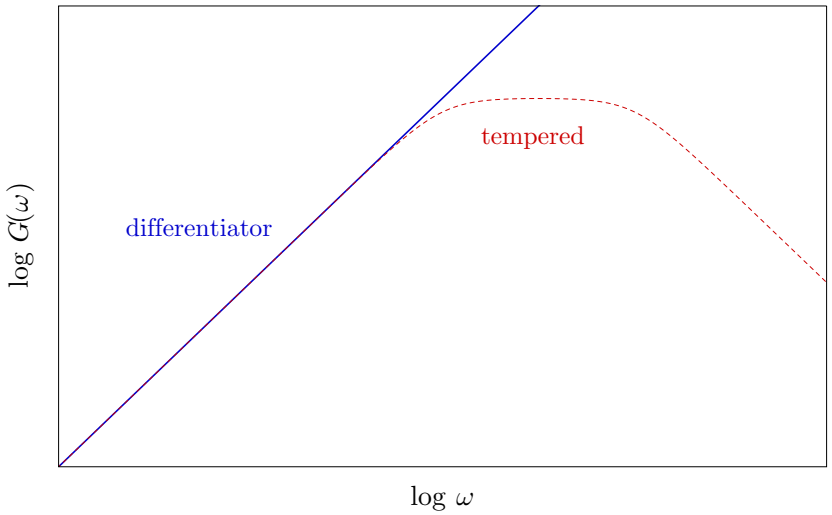
Like the differentiator, this amplifier gives the integral with an overall minus sign. It also depends on the initial output state.

7.4.3 Differentiator Issues

From Eq. (7.20), the main problem with the differentiator is that it has a gain that increases as ω, so the gain becomes arbitrarily large for large frequencies. This causes potential problems in two ways. First, differentiators suffer from bad high-frequency noise, and second, the high gain can possibly de-stabilize the amplifier. The solution is to add an extra, small, parallel capacitance in the feedback look, and an extra, small, series resistance in the input.



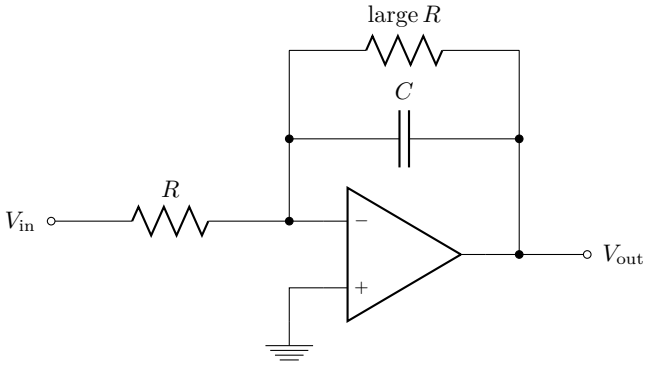
How do we think about this? First of all, the input network crosses over from capacitive to resistive at an input RC frequency. Above this frequency, the resistance dominates, and in combination with the feedback resistor, the amplifier acts like an inverting amplifier, where the gain is flat with frequency. The feedback network also defines a second RC frequency; above this frequency, the capacitor bypasses the resistor. Then the capacitor, in combination with the input resistor, makes the op-amp behave as an integrator, where the gain decreases with frequency. The net effect is shown schematically in the gain plot below.



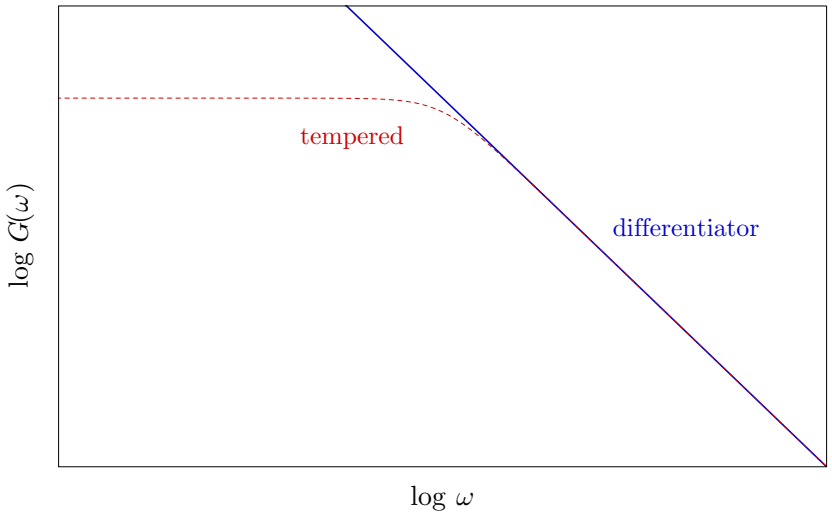
The solid line is the differentiator response, and the dashed line shows the "tempered" differentiator, including the effects of the input resistor (when the gain flattens) and the feedback capacitor (when the gain "rolls off," or decreases).

7.4.4 Integrator Issues

The integrator has the *opposite* problem: from Eq. (7.22), the gain scales as ω^{-1} over all frequencies, so the gain at dc diverges. This essentially means that the integrator has no "natural" dc level, and any dc input (even a spurious dc input) will eventually rail the op-amp. The fix for this is to put a large resistance in parallel with the feedback capacitor, as shown below.



The feedback network then defines an RC frequency (a small frequency, since the resistance is large), below which the resistor dominates the feedback impedance, and the op-amp acts as an inverting amplifier. This levels the low-frequency gain to some finite value, as illustrated schematically below.



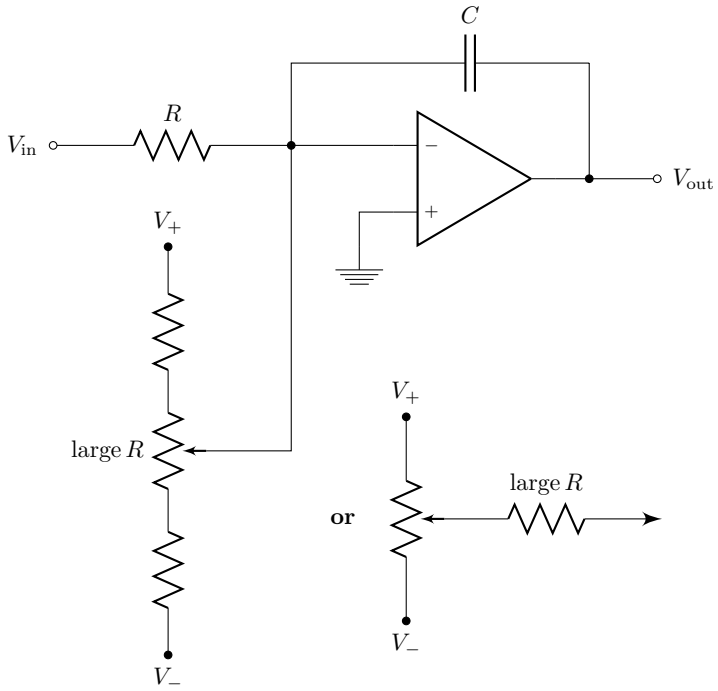
This shows the normal −6-dB scaling of the integrator as the solid line, with the dashed line showing the "tempered" behavior with the feedback resistor rolling off the low-frequency gain.

7.4.5 Sources of Integrator Error

We have mentioned that the basic integrator is very sensitive to spurious dc inputs, due to the divergent dc gain. What are these spurious dc offsets? There are two main sources for op-amps: input bias current, and input offset voltage.

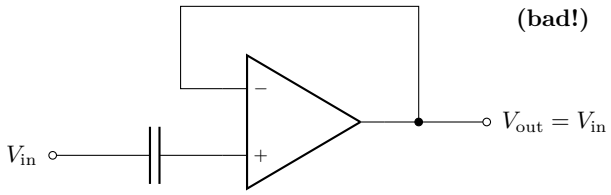
7.4.5.1 Input Bias Current

In the ideal op-amp we stated the golden rule that no current flows into the inputs. However, this isn't *quite* true. The idea behind the *input bias current* is that in fact a small current flows into (or out of) the inputs, which makes sense, as the op-amp inputs drive internal transistors, which either require current to work or allow a bit of leakage current to flow. For BJT-input op-amps, the input bias current is $\sim 10\,\text{nA}$, while for JFET-input op-amps, the input bias current is $\sim 10\,\text{pA}$. For example, the precision OPA602C with JFET inputs has a 1-pA input bias current (compared to the 741C with BJT inputs at 500 nA). Thus, JFET-input op-amps are the clear winner in this regard. To some extent, it is possible to compensate for the input bias current by injecting a small, adjustable current at the inverting input.

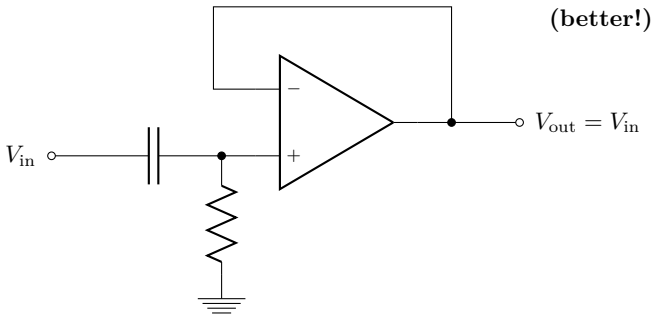


Here, $V_\pm$ are the power supplies. The idea here is essentially the same as for the summing amplifier: the adjustable voltage and resistance (or Thévenin resistance of the potentiometer) causes a current to flow—a small one, if the resistance is large. This can be adjusted to cancel the input bias current, for example by adjusting it until the output is stationary when $V_{\text{in}} = 0$. However, this isn't perfect: the input bias current depends on temperature, for example, so compensating at one temperature doesn't guarantee compensation everywhere. If this is a concern, it is far better to start with a good op-amp, rather than try to "fix" a crappier op-amp.

Another side effect of the input bias current is that the inputs need some dc path to ground. So, for example, it would be a bad idea to build an ac-coupled follower like this:



The inverting input is fine, as the path to "ground" is via the output. However, the noninverting input has no path to ground, and the input bias current will charge the capacitors until the inputs go out of range with respect to the power supplies, causing real problems. The fix is to use an input high-pass network, as shown below.



Now, the resistor supplies the path to ground. There is some offset voltage error given by the product of the bias current and the resistance, but this can be made small (in the previous circuit, this error was huge because the impedance to ground was effectively very large).

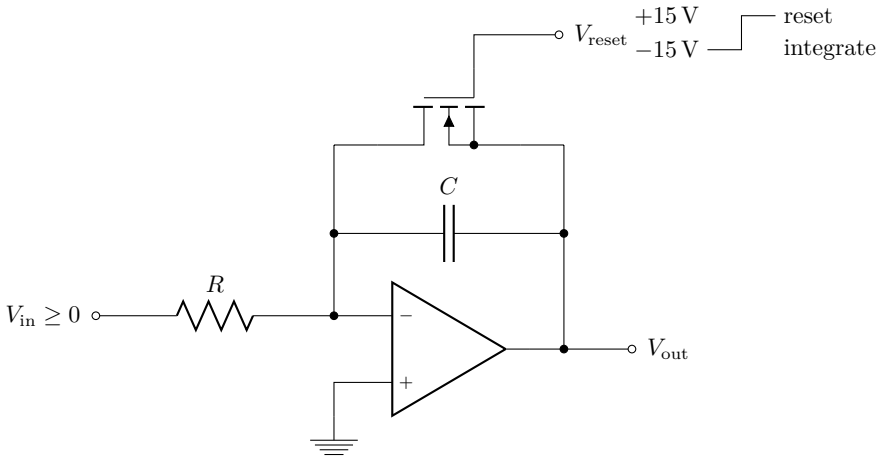
7.4.5.2 Input Offset Voltage

The other main effect is the **input offset voltage**. For an ideal op-amp, the output is zero when the inputs are exactly equal. But for a *real* op-amp, the output is zero when the inputs are *almost* the same, or said another way, when there is a small "error" voltage δV between the inputs. This is the input offset voltage, and is due to manufacturing variation when producing op-amp devices. Typically, the input offset voltage ranges from $\sim 10\,\mu$V to a few mV, with BJT inputs faring better than FET inputs. For example, the FET-input, precision OPA602C has 0.1 mV typical input offset voltage, and 0.25 mV max (compare to the 741C, which is 2 mV typical; this is not as much worse, when compared to the bias current). In the integrator, the net effect is that a zero V_{in} causes integration (the input should be set to the input offset voltage for no integration to occur). The compensation circuit above can also compensate for this effect, because it is equivalent to summing another input voltage with V_{in}. Most op-amps also have pin connections for a potentiometer to allow nulling of the input offset voltage (typically, for a single op-amp in a dual-inline package, a trim-pot is connected across pins 1 and 5). Again, while this can be trimmed, the drift will be of the same order as the uncompensated error, so in critical applications, it's better to choose an op-amp with a low offset voltage, rather than try to correct for the offset voltage of a "bad" op-amp.

7.4.6 Integrator Applications

The integrator is a widely useful circuit. One example is in feedback-control circuits (circuit to generate a stable voltage, current, temperature, etc.)—it turns out integration is useful in obtaining stable operation, a point to which we will return later.

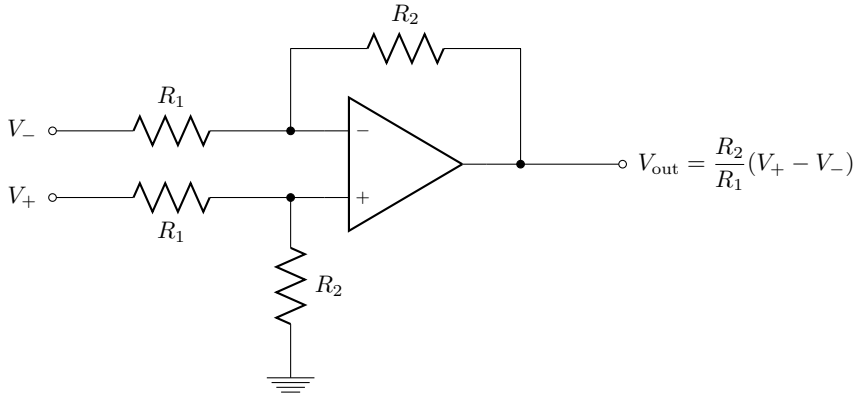
Literal integration of signals is also a useful task. For example, suppose we have an optical-pulse signal from a laser pulse on a photodetector. If we integrate the signal, we can get the pulse energy (or pulse **fluence**). In this case, it is useful to be able to reset the integrator just before we expect to receive each pulse, and the circuit below takes care of this.



The MOSFET here acts as an analog switch, that dumps the capacitor charge when it is necessary to reset the integrator. Note that we are assuming a positive input voltage, and thus a negative output voltage. If the output voltage may have either sign, a second, reversed MOSFET may be necessary to prevent dumping the capacitor charge during integration.

7.5 Instrumentation Amplifiers

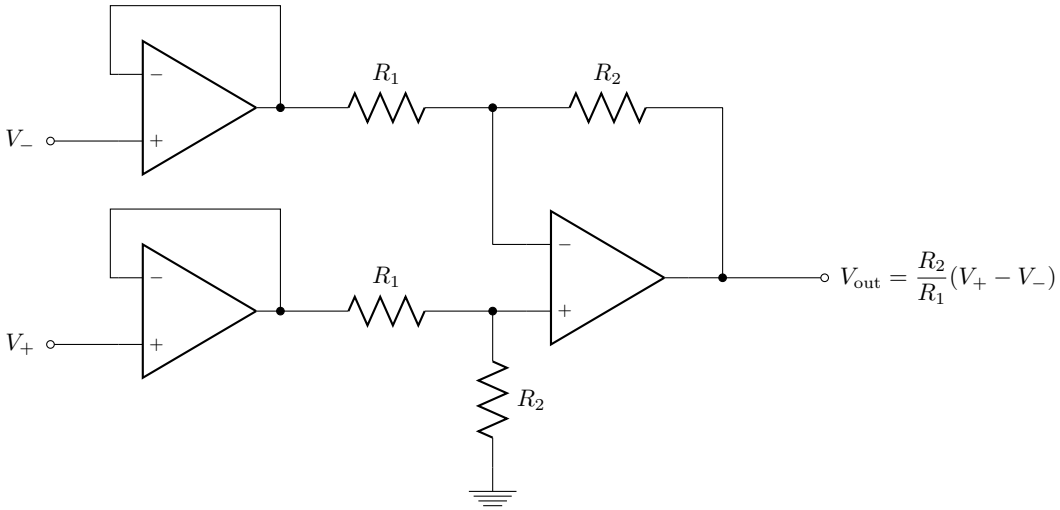
Recall the op-amp differential amplifier, shown below, from Section 7.3.5.



This circuit works fine, but has some disadvantages:

- This circuit does not have high input impedances, especially for large gains. Specifically, you should convince yourself that the noninverting input has an input impedance $R_1 + R_2$, while the inverting input has an input impedance R_1. But, for example, if $R_1 = 1\,\text{k}\Omega$ and $R_2 = 10\,\text{k}\Omega$, so that $G = 10$, the input impedance is at worst $1\,\text{k}\Omega$.
- The circuit requires accurately matched resistor pairs to achieve a high CMRR. To obtain a CMRR of 80 dB, the resistors must have a tolerance of around 0.01% at $G = 1$; resistors this accurate are typically wirewound, but these don't work well at high frequencies.
- Any source impedances add to the input R_1's. That is, the sources must act as ideal voltage sources, otherwise the gain and CMRR may be affected. For example, if $R_1 = 1\,\text{k}\Omega$, the source impedance must be $10\,\Omega$ or less for 1% gain accuracy. Even worse, the source impedances must be *matched* to ensure good CMRR.

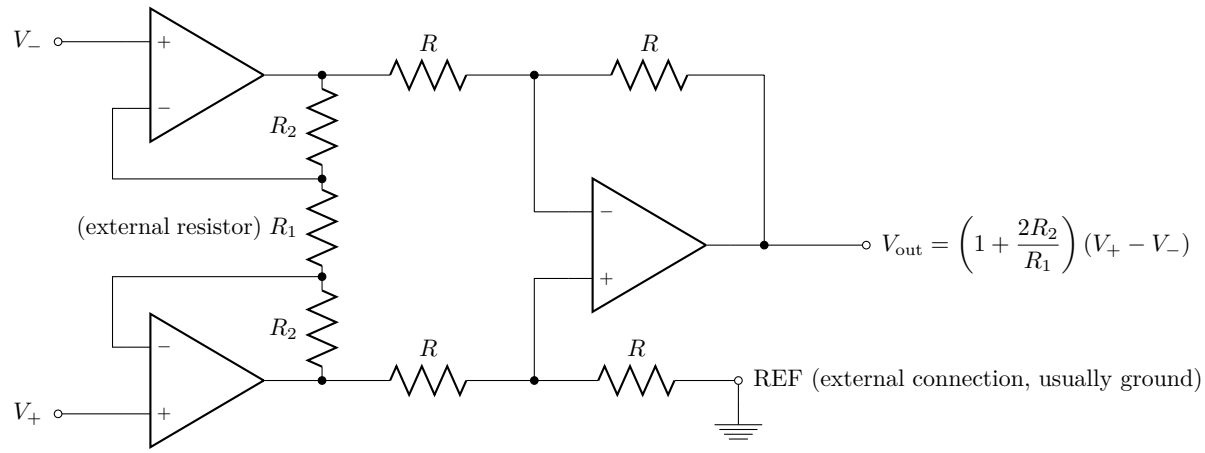
The main solution to these problems, especially that of source impedances interacting with the resistances in the op-amp circuit, is to simply buffer the inputs. A differential amplifier with buffered (high-impedance) inputs is called an **instrumentation amplifier**, and the basic circuit is shown below.



The point is that this entire circuit should come in a single package, with laser-trimmed (matched) resistor networks, for good performance and to make life easy. This also guarantees that any errors in the matched resistor pairs due to temperature drifts is kept to a minimum.

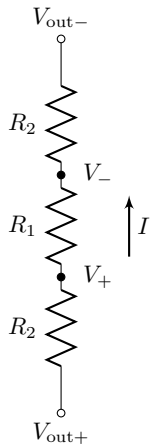
7.5.1 "Classic" Instrumentation Amplifier

What you *usually* find packaged as an instrumentation amplifier is actually a bit different than the above circuit. The inputs are still buffered for high impedance, but the input buffers are hooked up in a resistor chain as shown below, and the resulting difference is computed by a unity-gain differential amplifier.



The main advantage here is that the gain can be set by changing only *one* resistor, here R_1. This resistor is usually not included in the package, but rather the package has pins for an externally connected resistor for a user-settable gain. (You should be able to see that when R_1 is omitted, the input amplifiers reduce to buffers, and the output is just the difference of the inputs.) Another feature to note is that the ground connection of the differential amplifier is usually given as a "reference" (REF) pin on the package. This allows the subtraction to be referenced to another voltage besides ground, which is sometimes convenient. We will give an example below in Section 7.5.2.3.

To analyze the circuit above, let's focus on just the first resistor chain.



We have drawn in the voltages, assuming the op amps enforce the equality of their inputs, and we are labeling the outputs of the two input buffers as $V_{\text{out}\pm}$. No current flows into or out of the buffer-op-amp inputs, so we will assume a current I flows from $V_{\text{out}+}$ to $V_{\text{out}-}$. Using Ohm's law across R_1,

$$I = \frac{V_+ - V_-}{R_1}. \tag{7.26}$$

Then the voltage drop across the top resistor in the chain gives

$$V_{\text{out}-} = V_- IR_2 = V_- - \frac{R_2}{R_1}(V_+ - V_-), \tag{7.27}$$

and across the bottom resistor in the chain,

$$V_{\text{out}+} = V + + IR_2 = V_+ + \frac{R_2}{R_1}(V_+ - V_-). \tag{7.28}$$

Now the differential amplifier takes the difference between $V_{\text{out}+}$ and $V_{\text{out}-}$, so

$$V_{\text{out}} = V_{\text{out}+} - V_{\text{out}-} = (V_+ - V_-) + \frac{2R_2}{R_1}(V_+ - V_-), \tag{7.29}$$

or

$$V_{\text{out}} = \left(1 + \frac{2R_2}{R_1}\right)(V_+ - V_-). \tag{7.30}$$
(instrumentation-amplifier output)

As a gain, this reads

$$G = \left(1 + \frac{2R_2}{R_1}\right). \tag{7.31}$$
(instrumentation-amplifier gain)

Thus we see again if R_1 is omitted ($R_1 = \infty$), the gain is unity, while other values can serve to increase the gain above unity.

A good example of an instrumentation amplifier is the INA128 from Burr-Brown, one of a family of "INAXXX" instrumentation amplifiers. In the INA128, the internal R_2 resistors are 50 kΩ, so the external resistor R_{G} (i.e., R_1), sets the gain via

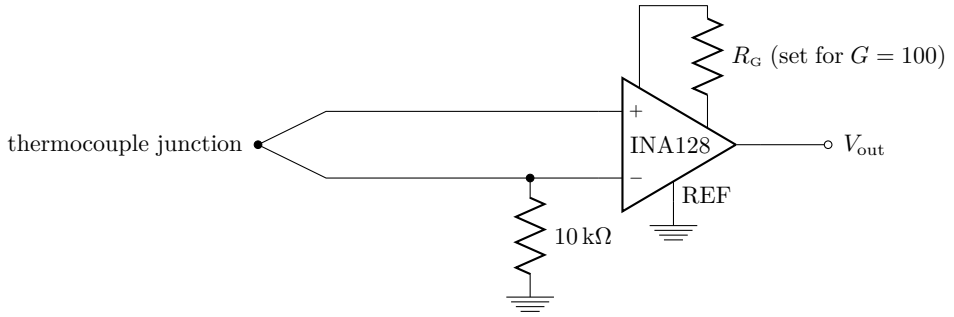
$$G = 1 + \frac{50\,\text{k}\Omega}{R_{\text{G}}}. \tag{7.32}$$

The CMRR is 86 dB at unity gain ($R_{\text{G}} = \infty$), and 125 dB at $G = 100$. The input impedance is $\sim 10^{10}\,\Omega$, and the input bias current is 2 nA. These cost ∼$8 each, depending on the grade (quality) and package.

7.5.2 Instrumentation-Amplifier Applications

7.5.2.1 Thermocouple Amplifier

The instrumentation amplifier turns out to be really useful in a number of applications, particularly in amplifying high-source-impedance sensors that require high gain. On example, shown below, is a simplistic thermocouple amplifier.

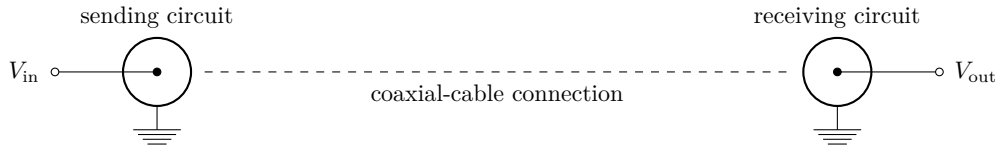


A **thermocouple** is a junction of two dissimilar metals that develops a voltage that is related (in an nonlinear way) to the temperature. Typical signal levels from thermocouples range from ∼10 μV to 100 μV, so significant gain is useful here. Note that this circuit is simplistic in the sense that it is only useful for measuring *relative* temperature changes. Absolute calibration requires computing the difference between two thermocouples, one held at a known reference temperature (e.g., an ice bath). This is called *cold-junction compensation*, and there are special amplifiers that can emulate the cold junction electronically—a good example is the AD594.

One other thing to note is the 10-kΩ resistor, which is necessary for the circuit to function. Why?

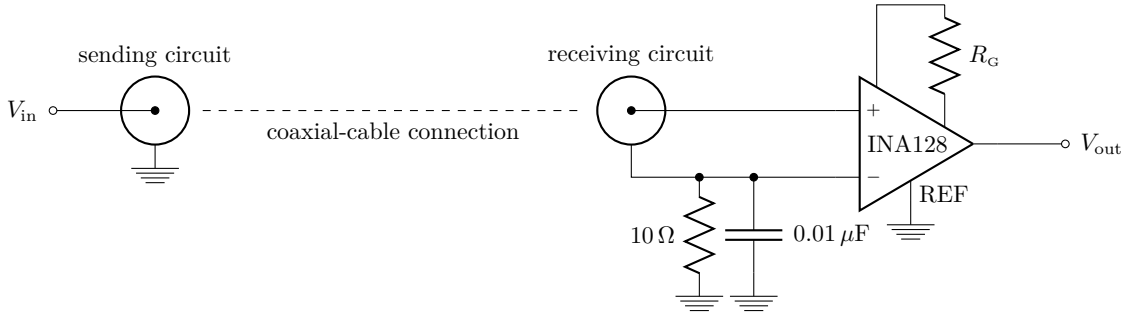
7.5.2.2 Differential Transmission for Noise Rejection

Another good application of the instrumentation amplifier is as a receiving input for a signal that is transferred between two instruments. For low noise, the signal is usually sent via **coaxial cable**—a center conductor shielded by a cylindrical outer conductor, which is usually used as a ground connection, so the grounded jacket protects the center (signal) conductor from external interference. The simplest way to send a signal between instruments is just to connect it as shown below—use a jack on either instrument, with the coaxial cable in between, the outer conductor grounded on either end.



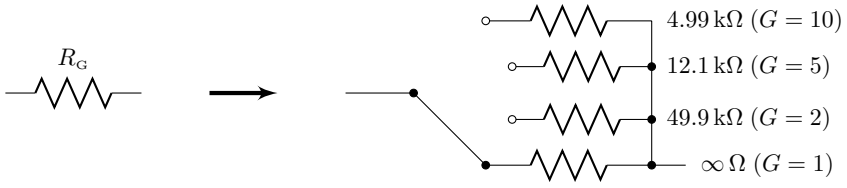
This is a bad idea, though, because it introduces a **ground loop**: both instruments are generally grounded via their power supplies, and the grounds are *also* connected via the coaxial-cable jacket. This means that the ground has a big loop for a path, and changing electromagnetic fields can induce and EMF in this loop. The symptoms are noise pickup, such as 60-Hz buzz in audio systems, or radio-frequency interference in wide-bandwidth circuits.

One solution is to power one instrument from a battery, so that it "floats" (i.e., there is no ground connection). This is not always convenient, but it is also possible to use an isolation transformer to break the ground connection via the power supply. (Isolation transformers for this purpose are commercially available, but usually require a safety ground connection to be defeated before it truly provides ground isolation.) However, having instruments grounded is otherwise desirable for safety and noise immunity, so another solution is to use an instrumentation amplifier as a differential receiver on the receiving instrument, as shown below.



This also breaks the ground loop, and any induced interference, which is common to the ground and signal conductors in the cable, will be cancelled in the subtraction. (This also works for twisted-pair cable, in place of the coaxial cable, as is usually used in Ethernet networks.) Note that the "ground" conductor on the received is still tied to ground via a resistor and capacitor. This prevent large induced input swings, in order to protect the amplifier (and remember we need the resistor to provide a dc path to ground).

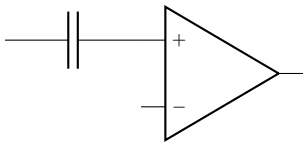
Another useful trick is to use the instrumentation amplifier to provide variable gain on the input, which is easy to accomplish by replacing R_G by a switch-selectable array of gain resistors (e.g., using a rotary switch to select gain), as shown below.



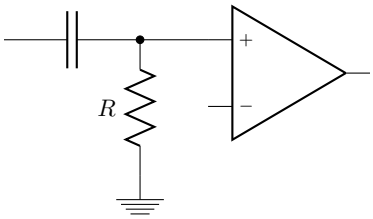
A logarithmic spacing of gains (1, 2, 5, 10, 20, 50, 100, ...) covers a wide gain range in a useful way.

7.5.2.3 AC-Coupled Inputs with High Impedance

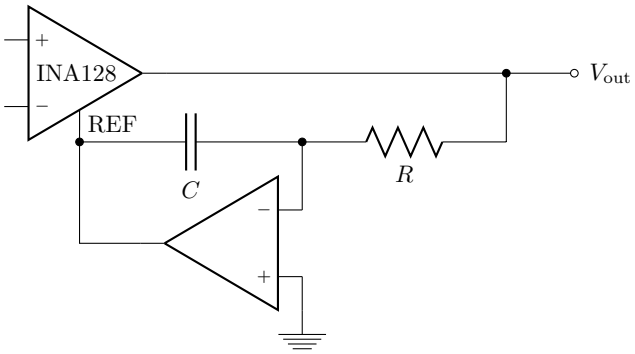
Sometimes it is useful to have an ac-coupled input, but with high input impedance. For example, we may have a sensor with a weak signal and high source impedance that requires large amplification, but if it has a large dc bias, a large gain would take the dc bias out of range. If we just use a capacitor to block dc on the input, this is bad, remember, because we need a dc path to ground on each input.



An improvement is to introduce a resistor, which makes a high-pass filter. The resistor prevents problems from input bias currents.



However, in the pass band, the input impedance is limited to R, which will not be nearly as good as a "bare" input on a decent op-amp. A clever solution is to use an integrator to feed back the output of an instrumentation amplifier to the reference input.



Assume for simplicity that the instrumentation amplifier is set for unity gain, so

$$V_{\text{out}} = V_+ - V_- + V_{\text{REF}}. \tag{7.33}$$

That is, the reference voltage is the reference for ground, so the amplifier sets $V_{\text{out}} - V_{\text{REF}}$ to the difference $V_+ - V_-$. Note that the only steady state occurs when the integrator is at steady state, which is when the integrator input (V_{out}) is zero. If $V_{\text{out}} \neq 0$, the integrator will build up a voltage until it cancels the steady output voltage on a time scale RC (with negative feedback due to the inverting nature of the integrator). High-frequency signals are not affected by the integrator, because the integrator has no time to "catch up" to the rapidly changing output (which is saying that the integrator gain is suppressed at high frequencies as $1/\omega$). To make this input switchable between ac and dc, a switch can short across the capacitor to change the instrumentation amplifier to dc mode.

7.6 Practical Considerations

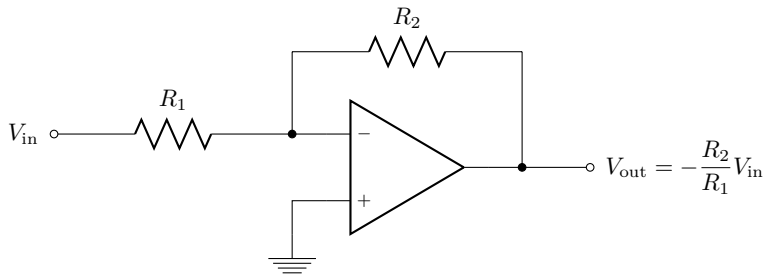
In this section, we will deal some more with some deviations of op-amps from their ideal behavior, and a few common tricks to mitigate these effects.

7.6.1 Input-Bias Currents and Precision Amplifiers

Before, when we were dealing with integrating amplifiers (Section 7.4.4), we talked about input-bias currents and how they can cause problems by charging up the integrating capacitor, even with a zero input voltage. We also talked about how it is critical to have a dc path to ground for each input to prevent similar charging problems. However, input bias currents can *still* cause problems (albeit usually less serious) in "regular" op-amp circuits like inverting and noninverting amplifiers.

7.6.1.1 Inverting Amplifier

For example, let's return to the inverting amplifier (Section 7.3.2).



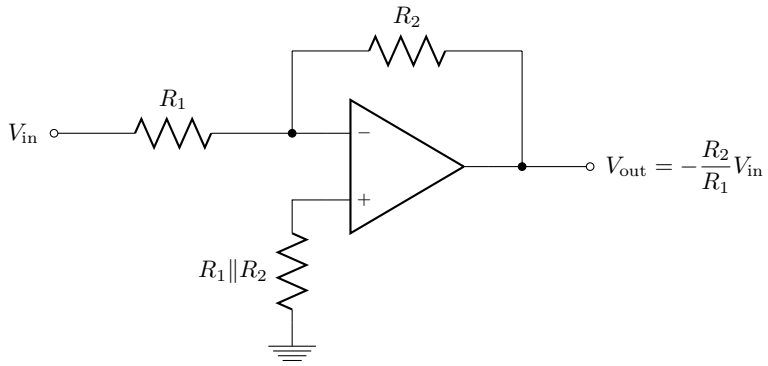
As a numerical example, consider a 741C (a crappy op-amp that will make such problems obvious), with a worst-case input bias current of $0.5\,\mu$A. If we take $R_1 = 10\,\text{k}\Omega$ and $R_2 = 1\,\text{M}\Omega$ (for $G = -100$), then the inverting input sees a Thévenin-equivalent input resistance of $R_{\text{Th}} = R_1 \| R_2 \approx 10\,\text{k}\Omega$. Then the input bias current leads to a bias voltage of $(0.5\,\mu\text{A})(10\,\text{k}\Omega) = 5\,\text{mV}$ at the input. With a gain of (-100), this leads to an output bias voltage of 5 V worst-case, which is pretty bad! (That is, for $V_{\text{in}} = 0$, it is permissible according to the 741C's specs to have $|V_{\text{out}}|$ as much as 5 V.)

How do we get around this? There are a few approaches:

- Since this is a dc-bias issue, this is no problem for *ac* circuits—just make sure to design for zero dc gain in the circuit (using a blocking capacitor, for instance).
- The main problem in the example was an underwhelming performer of an op-amp. Of course, we can do better, and recall that FET-input op-amps are superior to BJT-input op-amps in terms of bias currents. Precision BJT-input op-amps are a good option, too. For example, the FET-input LF411 has $I_{\text{bias}} \approx 0.2\,\text{nA}$, so the output error is only 0.2 mV in the above example.
- For BJT-input op-amps, or in high-precision circuits, we can also reduce errors by balancing input impedances, which requires a few examples.

7.6.1.2 Balanced Input-Impedances: Inverting Amplifier

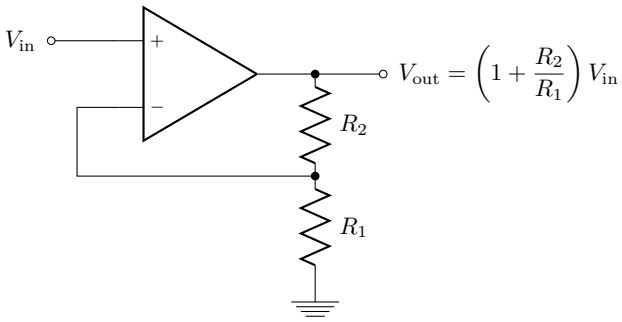
To balance the input impedances of the inverting amplifier, recall that the problem came from the inverting input "seeing" an effective source impedance of $R_1 \| R_2$. We can simply insert an equivalent source impedance on the *noninverting* input as shown below.



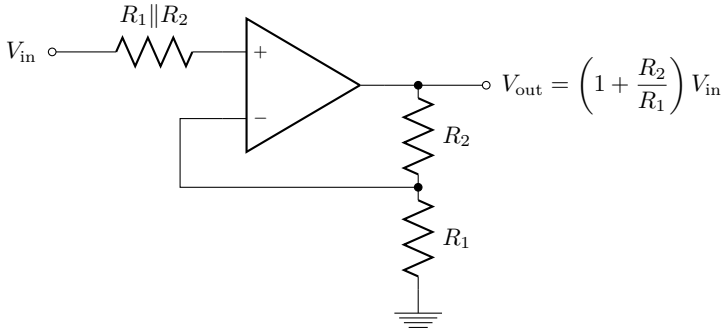
This resistor does nothing for an ideal op-amp, because no current flows through the resistor (and thus the resistor drops no voltage), but for a *real* op-amp, this trick reduces bias errors. Of course, all this assumes V_{in} acts like an ideal voltage source (i.e., it has a source impedance much smaller than R_1), otherwise the source impedance of V_{in} must also enter into the compensation scheme.

7.6.1.3 Balanced Input-Impedances: Noninverting Amplifier

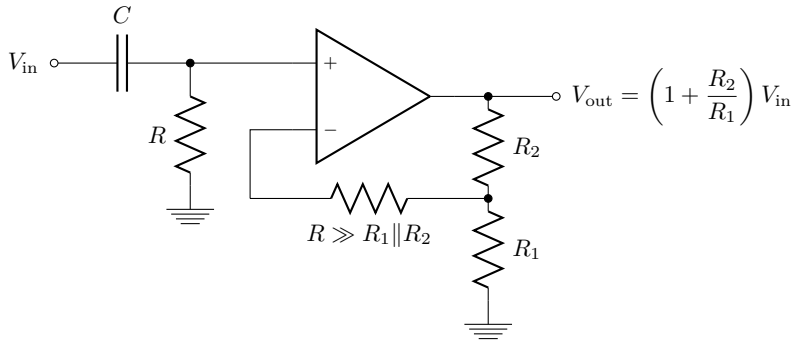
Similar balancing tricks are possible for noninverting amplifiers. Recall the basic noninverting amplifier (Section 7.3.3), shown below.



Here, the inverting input sees a source impedance of $R_1 \| R_2$, due to the voltage divider. Thus, it is again a simple matter to insert an equivalent source impedance for the other input.



Balancing input impedances is a bit trickier for ac-coupled amplifiers. A good example is shown below.



A high-pass filter is used on the input to block the dc component (remember the resistor R is necessary to provide a dc path to ground for the input bias current). We need to balance impedance at dc, so the noninverting input sees a source impedance of R. An easy way to ensure the same impedance for the inverting input is to insert another resistor R in the feedback loop, and simply ensure $R \gg R_1 \| R_2$, so we don't have to worry about R_1 and R_2 when setting the resistances.

7.6.1.4 Input Offset Currents

In addition to input bias current, there is **input offset current**, which is basically the difference between the input bias currents for the two inputs of an op-amp. While input bias current is more or less intrinsic to an op-amp's design, input offset current is due to manufacturing asymmetry. Thus, even with balanced inputs, there will be some bias signal, but much smaller than in the unbalanced case. For example, the 741C has an input offset current of 200 nA (compared to 0.5 μA for input bias current), and the LF411 has an input offset current of 0.1 nA (compared to 0.2 nA for input bias current).

7.6.1.5 Common-Mode Rejection Ratio

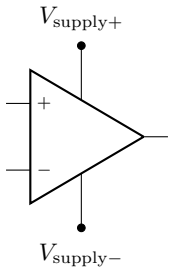
At this point, we can return to another source of error, the **common-mode rejection ratio (CMRR)**, which we introduced in Section 4.10.3 for the transistor differential amplifier, and we discussed it again for instrumentation amplifiers in Section 7.5. The typical CMRR range of op-amps is around 50–125 dB. But now look back at the op-amp inverting and noninverting amplifiers above. The *inverting* amplifier is *insensitive* to bad CMRR, compared to the *noninverting* amplifier, which is relatively *sensitive* to common-mode signals. Why is this? (What are the op-amp input voltages in each case?)

So while the noninverting amplifier looks good because of its very-high-impedance input, it is not quite as precise as the inverting amplifier. The latter is better in high-precision applications.

7.6.2 Power Supplies

So far, we haven't talked so much about the power-supply connections of op-amps. We have talked about the rule that no current flows in or out of the inputs (or at least there is only a very small current). Clearly, a current must flow in or out of the output in order for interesting things to happen. This output current must come from somewhere, and that is where the power-supply connections come in.

Op-amps have two supply connections, and they are often powered from **split supplies** of $\pm 15\,V$. More generally we will call these supply voltages $V_{\text{supply}\pm}$, and the explicit connections are shown below.



There are also **single-supply** op-amps, where $V_{\text{supply}-}$ can be ground. Usually these can *also* be powered by split supplies, but the difference is as follows. Op-amps involve transistors, so the output can usually only swing to within a volt or two of either power-supply rail. Single-supply op-amps are designed with outputs that can swing all the way to the negative rail (so the output can swing to ground if in a single-supply circuit).

7.6.2.1 Power-Supply Rejection

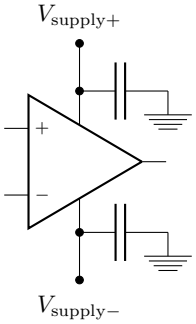
Ideally, the behavior of an op-amp is completely independent of the power supplies. That is, suppose the ideal op-amp output is 5 V, and the op-amp is powered from ±15 V. Then suppose the power supply changes to ±16 V. The output should be determined only by the circuit inputs and feedback network, and so shouldn't change at all, but in reality, it will change slightly, say to 5.1 V, to fabricate an example.

The (in)sensitivity of an op-amp to power-supply fluctuations is characterized by the **power-supply rejection ratio (PSRR)**. Ideally, the PSRR is very large, meaning the op-amp effectively "rejects" fluctuations in the power supply. The PSRR is defined with respect to the op-amp *inputs*, and is the ratio of the change in the power supply to the corresponding effect on the op-amp, referenced to a change at the input. (This accounts for the fact that power-supply fluctuations will have larger effects on the output for circuits with high gain.) As an example, suppose we have a PSRR of 120 dB, which is a ratio of 10^6 in voltage. Then a 1-V change in the power supply corresponds to a 1-μV change *at an input* to the op-amp. We must then get into the specific connections of the op-amp circuit (i.e., the gain) to determine the change in the output voltage. (So a unity-gain follower would also see a 1-μV change *at the output*.)

As a real example, the LF411 has a PSRR at dc of 100 dB typical, 80 dB minimum. The PSRR is better for the + supply than for the − supply, and the PSRR is worse at higher frequencies, dropping to ∼90 dB at 100 Hz, and ∼30 dB at 100 kHz.

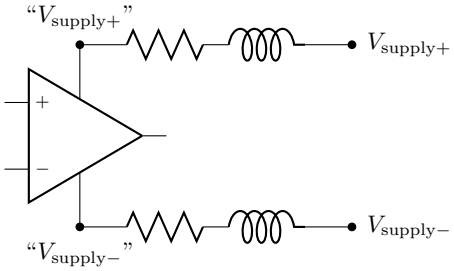
7.6.2.2 Power-Supply Bypass Capacitors

One common trick for improving the behavior of op-amp circuits is to use "bypass capacitors" on the power-supply leads of op-amps. The basic connection is shown schematically below.



The values of these bypass capacitors are not critical, but typically they would be 0.01-μF (monolithic) ceramic, 0.1-μF (monolithic) ceramic, or 1-μF tantalum (polarized) capacitors. These capacitors should also be placed physically as close as possible to the op-amp power-supply pins.

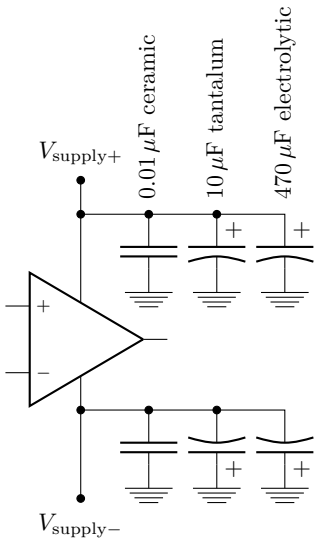
Why use bypass capacitors? In a circuit, for example, on a **printed circuit board (PCB)**, there may be long wires (or PCB traces) connecting the op-amp supply pins to the power supplies. Schematically this is shown below, where the wires have some intrinsic resistance and inductance.



The inductance is particularly problematic, as it means the path to the power supply has high impedance. The capacitor acts to short-circuit, or bypass, the inductance of the power-supply lead by providing a low-impedance path to ground at high frequencies. Otherwise, what can happen is as follows. A sudden change in the output of an op-amp (in response to an input change) implies a quick change in the power-supply currents. Inductance in the power-supply leads means that the voltage will drop if, for example, the op-amp is suddenly demanding more power-supply current. As we have seen, this can lead to output inaccuracies, because the op-amp PSRR is worse at high frequencies. In the worst case, the op-amp may even self-oscillate.

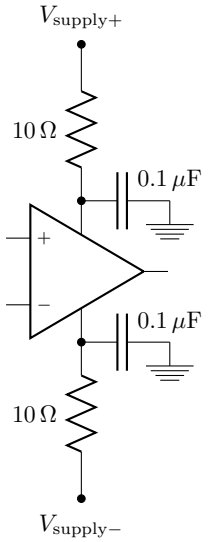
Another way to think of this is at the capacitor acts as a "charge reservoir" that tries to stabilize the power-supply voltages, temporarily supplying extra current as necessary when the op-amp demands it. Again, this is most effective at high frequencies, where the PSRR is bad anyway. Thus we also see the importance of using small, ceramic or tantalum capacitors (which have low inductance and respond well at high frequencies), and placing the capacitors very close to the op-amp (to bypass as much of the lead inductance as possible). On a PCB, each op-amp has its own local bypass capacitors, and the connection to ground should have very low impedance (e.g., to a **ground plane**, or a grounded copper layer that covers most of one side or layer of a PCB).

For critical applications, for example for a high-current amplifier, an even better approach is to use multiple, parallel bypass capacitors on each power-supply pin, as shown below.



Here the large 470-μF capacitor acts as a large charge reservoir, as appropriate for the high-current circuit, but only responds well at relatively low frequencies due to a high intrinsic inductance. Progressively smaller capacitors with lower inductance will help stabilize the power-supply voltage at higher frequencies. In this case, the smallest capacitors should be located closest to the op-amp, with the location of the large capacitors not so critical.

Another good approach for a low-current, or low-level, precision op-amp circuit is as shown below.



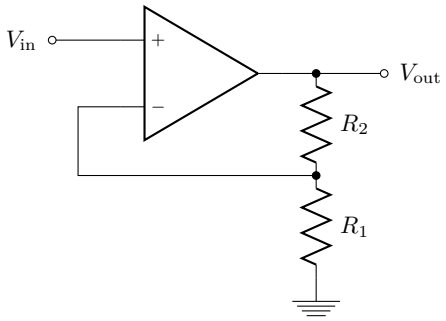
The idea is to add extra resistance to the power-supply lines, before the bypass capacitor. Obviously, this trick is limited to low-current circuits, so the voltage drop due to the 10-Ω resistor is small. This resistor enhances the effect of the bypass capacitor and gives improved isolation from power-supply fluctuations. This is an especially useful technique if high- and low-current amplifiers coexist in the same circuit, where feedback from the high-current to low-current amplifiers may cause the circuit to self-oscillate. Even better is to make sure the high-current and low-current ICs do not share the same power-supply lines (they should use separate lines, connected only at a point where the supply voltage is regulated, for example, by a 3-terminal regulator).

7.7 Finite-Gain Analysis

So far, we have assumed that the *open*-loop gain A of the op-amp is arbitrarily large. We have mainly made this assumption via the rule that the input voltages are the same in negative-feedback mode. In a *real* op-amp, this open-loop gain is high, but finite, ranging from about $\sim$50–146 dB. So now let's relax the assumption of infinite open-loop gain to (1) see what the effects are, and (2) to see why having a large gain is a good thing in an op-amp.

7.7.1 Noninverting Amplifier

As a first example, let's return to the noninverting-amplifier circuit of Section 7.3.3, shown below.



Recall that the result with infinite open-loop gain was

$$V_{\text{out}} = \left(1 + \frac{R_2}{R_1}\right) V_{\text{in}}. \tag{7.34}$$

But now, we will use the op-amp rule

$$V_{\text{out}} = A(V_{\text{in}+} - V_{\text{in}-}), \tag{7.35}$$

where A is the open-loop gain, instead of just assuming $V_{\text{in}+} = V_{\text{in}-}$. To simplify notation a bit, let's define the voltage-divider fraction

$$\eta := \frac{R_1}{R_1 + R_2}, \tag{7.36}$$

so that the divided output voltage ηV_{out} is fed back to the inverting input,

$$V_{\text{in}-} = \eta V_{\text{out}}. \tag{7.37}$$

Also, we have the input

$$V_{\text{in}} = V_{\text{in}+}, \tag{7.38}$$

so putting these equations into the op-amp rule (7.35), we have

$$V_{\text{out}} = A(V_{\text{in}} - \eta V_{\text{out}}). \tag{7.39}$$

Solving for the output voltage, we have

$$V_{\text{out}} = \frac{A}{1 + \eta A} V_{\text{in}},$$

(noninverting amplifier, finite open-loop gain) (7.40)

which defines the *closed*-loop gain

$$G = \frac{A}{1 + \eta A}.$$

(closed-loop gain, noninverting amplifier with finite open-loop gain) (7.41)

Note that in the limit $\eta A \gg 1$, this expression reduces to the original formula

$$G_\infty = \eta^{-1} = 1 + R_2/R_1, \tag{7.42}$$

so this analysis reproduces the ideal-op-amp limit.

7.7.1.1 Gain Limits and Error

Notice that $G \leq A$, so that the *open*-loop gain A limits the *closed*-loop gain G—negative feedback can only *reduce* the gain. Additionally, $G < G\infty$, so the *ideal* closed-loop gain always limits the *real* closed-loop gain. More specifically, assuming A to be large ($A \gg G_{\text{infty}}$),

$$G = \frac{A}{1 + \eta A} = \frac{\eta^{-1}}{1 + (\eta A)^{-1}} = \frac{G_\infty}{1 + G_\infty/A} \approx G_\infty \left(1 - \frac{G_\infty}{A}\right). \tag{7.43}$$

Then the fractional "error" in the gain is

$$\frac{\delta G}{G_\infty} \approx -\frac{G_\infty}{A}. \tag{7.44}$$

As an example, a decent op-amp has A of 100 dB, which corresponds to $A = 10^5$. For a $G_\infty = 10$ setup, the fractional error is

$$\frac{\delta G}{G_\infty} \approx -\frac{10}{10^5} = -10^{-4} = -0.01\%, \tag{7.45}$$

which is pretty small. For reasonably high values of A, this error is usually negligible compared to the error due to the feedback-resistor tolerances.

7.7.1.2 Insensitivity to Gain Variation

Another handy result that we obtain for large A is that if A is sufficiently large, then G is insensitive to variations in A. Starting with the closed-loop-gain expression above in terms of G_∞ and A,

$$G = \frac{G_\infty}{1 + G_\infty/A}, \tag{7.46}$$

then

$$\frac{\partial G}{\partial A} = -\frac{G_\infty}{(1 + G_\infty/A)^2}(-G_\infty/A^2) = \frac{G}{A^2}\frac{G_\infty}{(1 + G_\infty/A)} = \frac{G}{A}\frac{1}{(1 + A/G_\infty)}. \tag{7.47}$$

Then the variation δG in the closed-loop gain G is

$$\frac{\delta G}{G} = \frac{\partial G}{\partial A}\delta A, \tag{7.48}$$

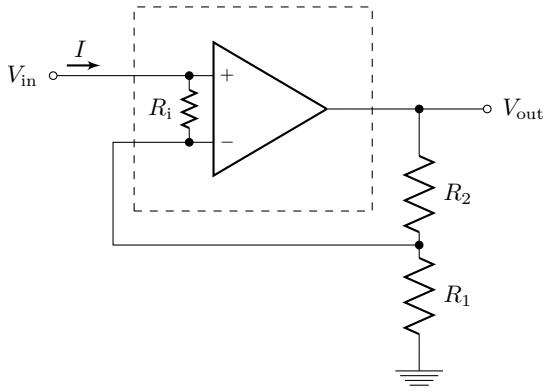
so

$$\frac{\delta G}{G} = \frac{1}{(1 + A/G_\infty)}\frac{\delta A}{A}, \tag{7.49}$$

so the fractional error $\delta G/G$ is much smaller than the fractional open-loop variation $\delta A/A$ by a factor of $1/(1 + A/G_\infty) \approx G_\infty/A$ for large A. This is a nice property because, for example, feedback with large A reduces variations of gain with frequency, for a flatter response in a closed-loop amplifier. It also reduces nonlinearity and distortion, which you can roughly think of as variations in gain with signal amplitude.

7.7.2 Feedback and Input Impedance

Negative feedback with large open-loop gain also helps quite a bit with input and output impedance. Going back to the noninverting amplifier, we can construct an explicit model for input impedance R_i as shown below.



The dashed line encompasses the "real" amplifier, which consists of an ideal op-amp and a resistor modeling the input impedance. As before, the closed-loop gain (7.41) is

$$G = \frac{A}{1 + \eta A}. \tag{7.50}$$

Since, $V_{\text{in}+} = V_\text{in}$, and $V_{\text{in}-} = \eta V_\text{out}$, we can take the input current I to be

$$I = \frac{V_{\text{in}+} - V_{\text{in}-}}{R_\text{i}} = \frac{V_\text{in} - \eta V_\text{out}}{R_\text{i}} = \frac{1 - \eta A/(1 + \eta A)}{R_\text{i}}V_\text{in} = \frac{V_\text{in}}{(1 + \eta A)R_\text{i}}. \tag{7.51}$$

Then the effective input impedance Z_in is

$$Z_\text{in} = (1 + \eta A)R_\text{i} = \left(1 + \frac{A}{G_\infty}\right)R_\text{i},$$

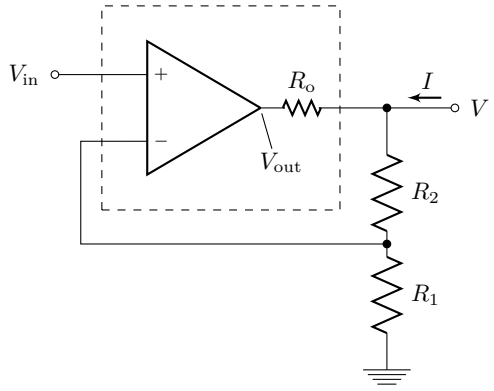
(input impedance, noninverting amplifier) (7.52)

which is much larger than the intrinsic input impedance R_i if there is large open-loop gain $A \gg G_\mathrm{infty}$.

For example, the modest 741C has $R_\mathrm{i} = 2\,\mathrm{M\Omega}$ typical, $300\,\mathrm{k\Omega}$ minimum, which is not great. The open-loop gain A is typically 2×10^5, minimum 1.5×10^4, which is not bad. If $G_\infty = 10$, then $Z_\mathrm{in} = 4 \times 10^{10}\,\Omega$ typical, or $5 \times 10^8\,\Omega$ minimum. These are pretty high input impedances, and they can be much higher with a precision op-amp.

7.7.3 Feedback and Output Impedance

To model the effects of feedback on the output impedance, we can again introduce an explicit model, including an output resistance R_o.



Again, the dashed box represents the "real" amplifier, with an ideal op-amp and the output resistor. We will call the output of the ideal amplifier V_out, while the "real" output is V. There is also a current I, which we define as flowing *into* the output. Then setting a null input $V_\mathrm{in} = 0$, we have $V_{\mathrm{in}+} = 0$, and now $V_{\mathrm{in}-} = \eta V$, so the op-amp rule (7.35) gives

$$V_\mathrm{out} = -\eta A V. \tag{7.53}$$

Then the current is

$$I = \frac{V - V_\mathrm{out}}{R_\mathrm{o}} = \frac{V(1 + \eta A)}{R_\mathrm{o}}. \tag{7.54}$$

Thus, the output impedance $Z_\mathrm{out} = V/I$ is

$$Z_\mathrm{out} = \frac{R_\mathrm{o}}{1 + \eta A} = \frac{R_\mathrm{o}}{1 + A/G_\infty}.$$

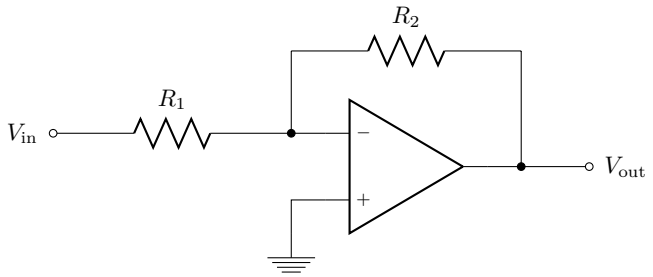
(output impedance, noninverting amplifier) (7.55)

This should be much smaller than the intrinsic output impedance R_o, provided we have large open-loop gain, $A \gg G_\infty$.

For example, the modest 741C has $R_\mathrm{o} = 75\,\Omega$. With a typical open-loop gain A of 2×10^5, and $G_\infty = 10$, then $Z_\mathrm{out} = (75\,\Omega)/(2 \times 10^4) \approx 4\,\mathrm{m\Omega}$, which is quite small.

7.7.4 Circuit Practice: Finite Gain in the Inverting Amplifier

For practice in dealing with finite op-amp gain, consider the noninverting amplifier, with finite open-loop gain A. (Again, it's best to do this before continuing, so we won't defer this until the end of the chapter.)



Show the following:
(a) The finite-A gain is

$$G = -\frac{(1-\eta)A}{1+\eta A}, \qquad \eta := \frac{R_1}{R_1+R_2}.$$

(closed-loop gain, inverting amplifier with finite open-loop gain) (7.56)

(b) Take the $A \longrightarrow \infty$ limit of Eq. (7.56) and show that

$$G_\infty = -\frac{R_2}{R_1}.$$

(closed-loop gain, inverting amplifier with infinite open-loop gain) (7.57)

(c) The input impedance is

$$Z_{\text{in}} = R_1 + \frac{R_2}{1+A}.$$

(input impedance, inverting amplifier with finite open-loop gain) (7.58)

(Ignore any intrinsic input impedance of the op-amp, which we will assume is much larger than R_1.)
(d) The output impedance is

$$Z_{\text{out}} = \frac{R_{\text{o}}}{1+\eta A},$$

(input impedance, inverting amplifier with finite open-loop gain) (7.59)

where R_{o} is the intrinsic output impedance of the op-amp, as in the noninverting case.

Solution.
(a) First, the noninverting input has $V_{\text{in}+} = 0$. The inverting input has a voltage determined by a voltage divider between V_{in} and V_{out}:

$$V_{\text{in}-} = \eta V_{\text{out}} + (1-\eta)V_{\text{in}}. \tag{7.60}$$

Remember $\eta = R_1/(R_1+R_2)$, so as a sanity check, $\eta \longrightarrow 1$, $V_{\text{in}-}$ becomes connected to V_{out}, and as $\eta \longrightarrow 0$, $V_{\text{in}-}$ becomes connected to V_{in}, which makes sense. Then using Eq. (7.35),

$$V_{\text{out}} = -\eta A V_{\text{out}} - (1-\eta)AV_{\text{in}}. \tag{7.61}$$

Solving for V_{out},

$$V_{\text{out}} = -\frac{(1-\eta)A}{(1+\eta A)}V_{\text{in}}. \tag{7.62}$$

This is the result we wanted, with G the coefficient of V_{in}.
(b) As $A \longrightarrow \infty$, $G = -(1-\eta)A/(1+\eta A) \longrightarrow -(1-\eta)/\eta = -R_2/R_1$.
(c) Suppose a current I flows into the V_{in} terminal. Then

$$I = \frac{V_{\text{in}} - V_{\text{in}-}}{R_1} = \frac{V_{\text{in}} - \eta V_{\text{out}} - (1-\eta)V_{\text{in}}}{R_1} = \frac{\eta(V_{\text{in}} - V_{\text{out}})}{R_1} = \frac{\eta[(1+\eta A)+(1-\eta)A]}{(1+\eta A)R_1}V_{\text{in}} = \frac{\eta(1+A)}{(1+\eta A)R_1}V_{\text{in}}, \tag{7.63}$$

where we used the solution (7.62). Then $Z_{\text{in}} = V_{\text{in}}/I$, so

$$Z_{\text{in}} = \frac{(1+\eta A)R_1}{\eta(1+A)} = \frac{R_1}{\eta(1+A)} + \frac{AR_1}{(1+A)} = \frac{R_1+R_2}{(1+A)} + \frac{AR_1}{(1+A)} = \frac{R_2}{(1+A)} + R_1. \tag{7.64}$$

Note that this reduces to R_1 as $A \longrightarrow \infty$.
(d) Here, we set $V_{\text{in}} = 0$ and call the output V, with a current I going into the output terminal. V_{out} is the output voltage of the ideal op-amp, before the intrinsic resistor R_{o}, as in the noninverting case. Then $V_{\text{in}-} = \eta V$, and

$$V_{\text{out}} = A(V_{\text{in}+} - V_{\text{in}-}) = -AV_{\text{in}-} = -\eta AV. \tag{7.65}$$

So the current is

$$I = \frac{V - V_{\text{out}}}{R_{\text{o}}} = \frac{1 + \eta A}{R_{\text{o}}} V, \tag{7.66}$$

and so the output impedance $Z_{\text{out}} = V/I$ is

$$Z_{\text{out}} = \frac{R_{\text{o}}}{1 + \eta A}. \tag{7.67}$$

Note that this decreases to zero as $A \longrightarrow \infty$.

7.8 Bandwidth

The **bandwidth** of an amplifier refers to the frequency range over which the response (gain) is reasonably flat. For electronic amplifiers, one characteristic is that the gain must fall off above some frequency—no amplifier can work at arbitrarily high frequencies.

Recall that the closed-loop gain G and the open-loop gain A are related, in that the latter bounds the former:

$$G \leq A. \tag{7.68}$$

Now let's consider the frequency dependence of the gain. In particular, the open-loop gain $A(\omega)$ typically has a "one-pole response," like that of a low-pass filter:

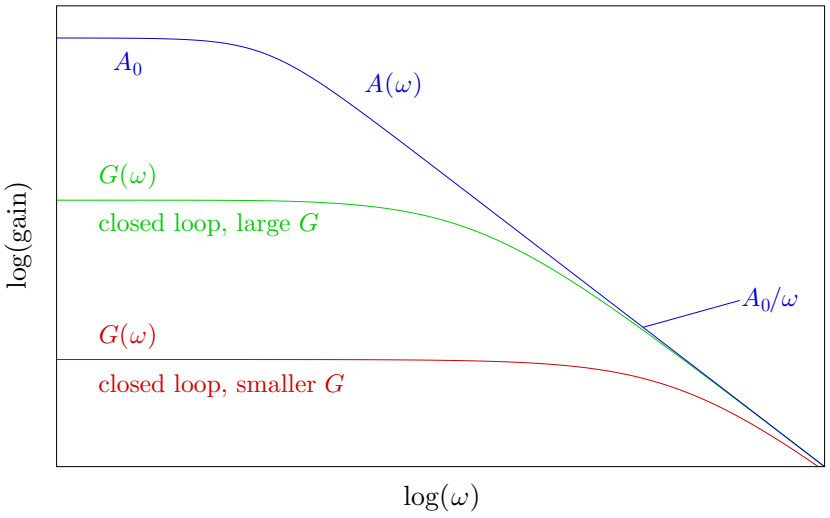
$$A(\omega) = \frac{A_0}{\sqrt{1 + (\omega/\omega_{3\,\text{dB}})^2}}. \tag{7.69}$$

Here, the cutoff frequency $\omega_{3\,\text{dB}} = 1/RC$ for a low-pass filter, where R and C are typically set by intrinsic transistor resistance and stray or added (internal) capacitance in the-op-amp. Note that asymptotically,

$$A(\omega) \sim \frac{1}{\omega} \tag{7.70}$$

for large ω, for a scaling of $-6\,\text{dB/octave}$.

Then the closed-loop gain is bounded by the open-loop gain, so that as the open-loop gain falls off, so does the closed-loop gain. This is illustrated schematically below for two different dc gains G and an open-loop gain $A(\omega)$.



Note that as the dc gain becomes smaller, the bandwidth (frequency range over which the gain is roughly constant) becomes wider. Since A cuts off as ω^{-1}, the closed-loop gain $G(\omega)$ meets $A(\omega)$ at a frequency that scales in the same way as ω^{-1}. That is, the bandwidth scales as $1/G_0$, where G_0 is the dc gain. Said differently, the product of the dc gain G_0 and the bandwidth is a constant, and this is often quoted as the **gain–bandwidth product (GBWP)**, or the **unity-gain bandwidth**. For example, the 741C has a GBWP of 1.5 MHz. Generally speaking, op-amps tend to be slow, especially at high gains, compared to discrete transistors.

7.8.1 Slew Rate

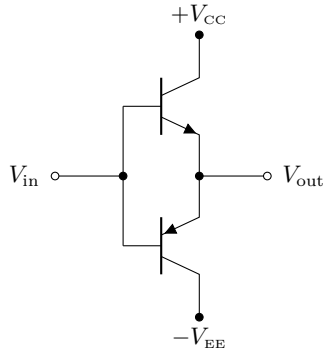
A concept closely related to bandwidth is the **slew rate**, which is the maximum rate of change of the output. Intuitively, this should be proportional to the GBWP, but this is somewhat more complicated because the same signal, but with different amplitudes, would involve different slew rates, even if they have the same frequency spectrum. So for rapidly changing signals, an op-amp with a particular slew rate may be able to follow the signal at low amplitudes, but it may be harder for the op-amp to follow the same signal at larger amplitudes.

As a concrete example, the 741C has a modest slew rate of $0.5\,\mathrm{V}/\mu\mathrm{s}$. Slew rates can be much high; for example, the BUF634 unity-gain buffer has a slew rate of $2000\,\mathrm{V}/\mu\mathrm{s}$.

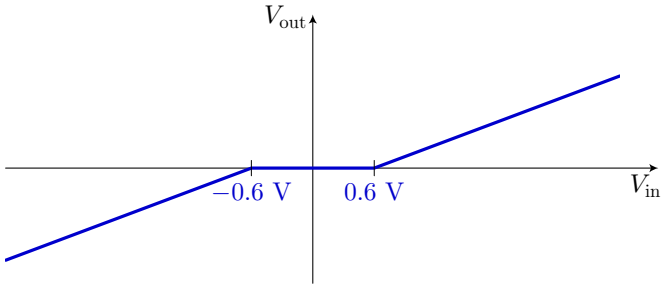
Generally speaking, the speed of an op-amp (either in terms of slew rate or GBWP) is controlled by the internal capacitance, which is usually fixed by an internal compensation capacitor, but also by intrinsic emitter resistance. Recalling that $r_\mathrm{e} \propto 1/I_\mathrm{C}$, generally speaking, a larger quiescent current (idling current) for an op-amp gives a higher slew rate or a wider GBWP. There is thus a trade-off between power and speed—some op-amps, like the OPA602, have a programmable quiescent current so the user can choose exactly where to make this trade-off.

7.8.1.1 Slew Rate and Power-Boosted Op-Amps

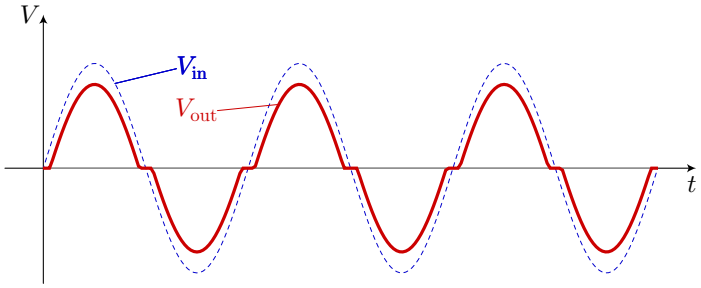
There are certain circuits where the slew rate of an op-amp is critical to its performance. One example is a "power-boosted" op-amp, where transistors are used to boost the output current capacity of an op-amp. The motivation for this circuit comes from the following "push-pull" current amplifier.



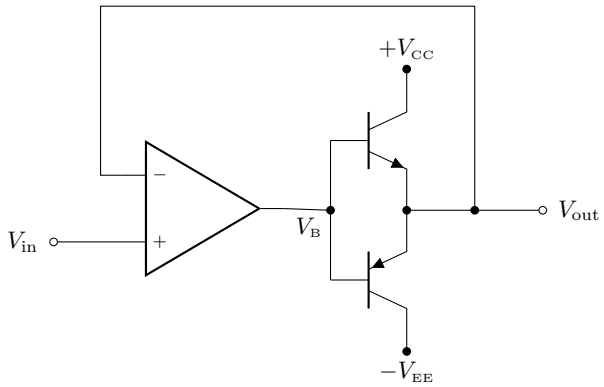
This is basically a pair of emitter followers. The potential advantage is operation with input signals of either polarity. The problem, though, is that one of the transistors will conduct, and the emitter (output) voltage must be a diode drop closer to zero than the base (input) voltage. That is, a graph of the output voltage responding to input voltage is schematically as in the graph below, if we assume the simple model that the base-emitter voltage drop is a constant 0.6 V (or less).



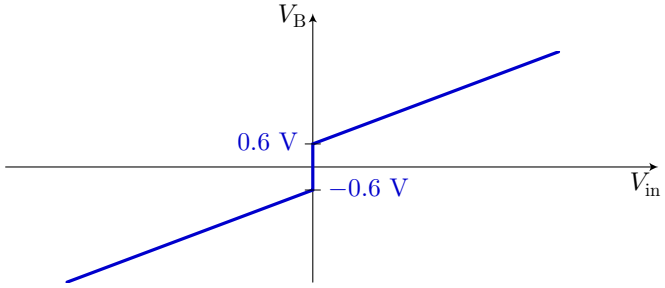
The problem in using this as an amplifier is that it leads to **crossover distortion**, because the base-emitter drop changes as the signal crosses through zero. An example is shown in the graph below of crossover distortion of an input sine wave.



One nice solution, at least in principle, is to use an op-amp, and enclose the push-pull transistor pair in the feedback loop of the op-amp, as in the circuit below.



This circuit acts as a unity-gain buffer with high current-driving capacity, because the op-amp does whatever it needs to do to ensure that V_{out} is the same as V_{in}. And to do this, it must "undo" the crossover distortion, so the base voltage V_{B} in this circuit must respond to the input as in the graph below.



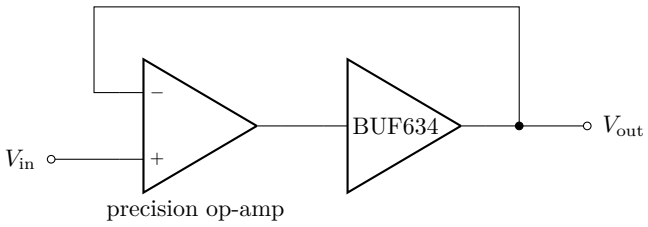
This response combined with the crossover distortion results in, in principle, a distortion-free output.

However, the problem with this conclusion is that it assumes that the op-amp has a long time to settle to the "correct" value. But with a rapidly changing input signal, the op-amp must jump discontinuously by

1.2 V when the input signal crosses through zero, which in practice can be problematic. This can lead to larger distortion and "glitching" with faster input signals.

A solution to this problem is to "bias" the transistors into conduction: basically, add 0.6 V to V_B at the npn's base, and -0.6 V to V_B at the pnp's base. Then with a 0-V input, both transistors are slightly conducting, but in opposition so their currents cancel. This wastes a bit of power at idle, but largely removes the problem with crossover distortion. The uncorrected amplifier is called a **class-B amplifier**, while the bias-corrected amplifier is called a **class-AB amplifier**.

Practically, it is complicated to design a bias-corrected circuit, especially to avoid thermal problems and proper selection and matching of bias voltages. For physicists, a simpler solution is to use a high-current buffer amplifier, where engineers have already taken care of the effort of biasing the push-pull pair. One example is the circuit below, which can handle 250-mA output signals via a BUF634 unity-gain buffer.



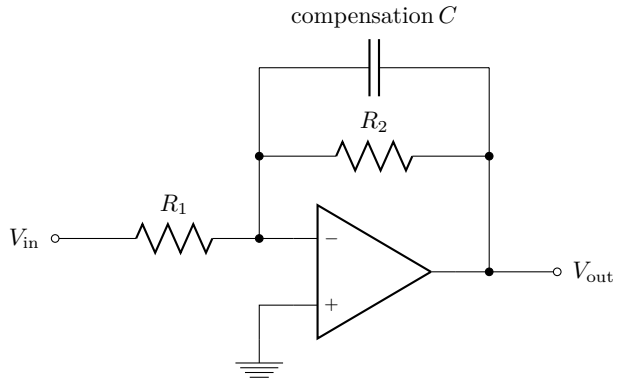
The idea is to use a precision op-amp, and take the feedback from the output of the BUF634. This way, we get the high-current capacity of the "slave" BUF634, combined with the precision of the "master" op-amp. One caveat, which we will explore in more depth, is that the buffer amplifier must have a much wider bandwidth than the master op-amp.

7.8.2 Stability and Compensation

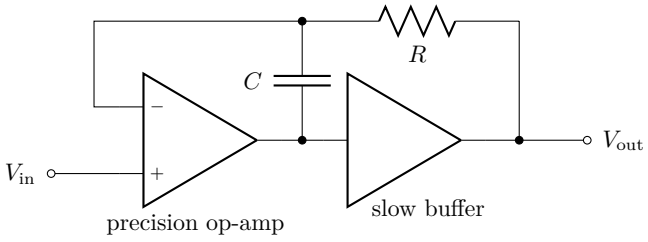
So far, we have talked about the frequency-dependent gain, but the frequency-dependent phase is also critical. As we have noted, for most purposes we can regard an op-amp as having a gain behavior similar to a low-pass filter.

Recall that for an RC filter, the relative phase of the output is 0° in the limit $\omega \longrightarrow 0$, and changes to a 90° phase lag as $\omega \longrightarrow \infty$. If we have multiple, cascaded filters, at high frequencies, intuitively we can think of having a 90° phase **per "pole"**, or roughly speaking, per RC pair. This can pose a problem for op-amp circuits that require negative feedback. Due to phase shifts and time delays in the feedback loop at high frequencies can add up to a 180° phase shift. However, negative feedback in combination with a phase shift of 180° (or in fact anything between 90° and 270°) is in fact *positive* feedback. This can turn into unstable behavior (oscillation) if the gain of the circuit exceeds unity in the frequency range where the feedback becomes positive feedback.

Thus comes the idea of **compensation**. Most op-amps include an internal capacitor to "roll off" the gain, and in particular to ensure that the gain is less than unity at frequencies where large phase shifts may cause problems. There are also **uncompensated op-amps**, which need an external capacitor or an appropriate reactive load to achieve stability. An example is the inverting amplifier below, with an explicit compensation capacitor to cut off the gain at high frequencies.



Another example is a variation on the BUF634 buffered op-amp circuit from the previous section. If the slave buffer amplifier is slower than the master op-amp, this may cause a problem because the phase shift due to the slower buffer may cause the master to become unstable or oscillate. A solution to use a slow buffer is the circuit below.



For small frequencies ($\omega \ll \omega_{3\,\text{dB}} = 1/RC$), the feedback comes from the buffer output, while at high frequencies ($\omega \gg \omega_{3\,\text{dB}} = 1/RC$), the feedback comes from the output of the master op-amp. This arrangement avoids problems with the phase shift and maintains stability of the amplifier.

7.8.2.1 Op-Amp Output and Capacitive Loads

To examine a problematic situation in a bit more detail, let's return to the bandwidth argument of Section 7.8, but now keep the complex phase in the open-loop gain:

$$\tilde{A}(\omega) = \frac{A_0}{1 - i\omega/\omega_{3\,\text{dB}}}. \tag{7.71}$$

Again, this response has the form of a low-pass filter [Eq. (2.44)], where the op-amp response "rolls off" due to a single capacitor. Then, taking the example of the noninverting amplifier of Section 7.7.3, the op-amp output impedance (7.55) becomes

$$Z_{\text{out}}(\omega) = \frac{R_{\text{o}}}{1 + A(\omega)/G_\infty} = \frac{R_{\text{o}}(1 - i\omega/\omega_{3\,\text{dB}})}{1 + A_0/G_\infty - i\omega/\omega_{3\,\text{dB}}}.$$

(output impedance, noninverting amplifier) (7.72)

Since typically $A_0 \gg G_\infty$, we can consider the intermediate range of frequencies between $\omega_{3\,\text{dB}}$ and $(A_0/G_\infty)\,\omega_{3\,\text{dB}}$, where the impedance reduces to

$$Z_{\text{out}}(\omega) = \frac{R_{\text{o}}(-i\omega/\omega_{3\,\text{dB}})}{A_0/G_\infty} = -i\omega\frac{R_{\text{o}}G_\infty}{\omega_{3\,\text{dB}}A_0} \qquad \left(1 \ll \frac{\omega}{\omega_{3\,\text{dB}}} \ll \frac{A_0}{G_\infty}\right). \tag{7.73}$$

This has the form of an inductive reactance [see Eq. (2.40)], with effective inductance

$$L_{\text{eff}} = \frac{R_{\text{o}}G_\infty}{\omega_{3\,\text{dB}}A_0}. \tag{7.74}$$

This inductive regime can span a wide range. With $\omega_{3\,\mathrm{dB}}/2\pi \sim 100\,\mathrm{Hz}$, $R_\mathrm{o} \sim 100\,\Omega$, $A_0 \sim 10^6$, and $G_\infty = 10$, this works out to a frequency range of $\sim$100 Hz to $\sim$10 MHz, with an effective inductance of $\sim$1.6 μH.

Outside this frequency band, the output impedance is simply resistive. For small frequencies, we have

$$Z_\mathrm{out}(\omega) = \frac{R_\mathrm{o}}{1 + A_0/G_\infty} \qquad \left(\frac{\omega}{\omega_{3\,\mathrm{dB}}} \ll 1 < \frac{A_0}{G_\infty}\right), \tag{7.75}$$

while for very large frequencies,

$$Z_\mathrm{out}(\omega) = R_\mathrm{o} \qquad \left(1 < \frac{A_0}{G_\infty} \ll \frac{\omega}{\omega_{3\,\mathrm{dB}}}\right). \tag{7.76}$$

Note that in practice, other stray capacitances will become important in the high-frequency range.

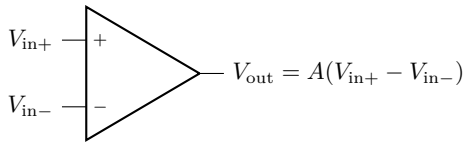
The inductive output of the op-amp in the regime of Eq. (7.73) can give rise to problems if the output of the op-amp drives a capacitive load. Then L_eff and the capacitance form a resonant circuit. This can lead to a resonant peak in the gain profile at frequency

$$\omega_0 = \frac{1}{\sqrt{L_\mathrm{eff} C}} = \sqrt{\frac{\omega_{3\,\mathrm{dB}} A_0}{R_\mathrm{o} C G_\infty}}. \tag{7.77}$$

For the same parameters, this would lead to a resonance at $\omega/2\pi \sim 100\,\mathrm{kHz}$ for a load capacitance $C = 1\,\mu\mathrm{F}$. If the effect of the resonance is sufficiently strong (i.e., sufficiently large Q), the circuit can become unstable and oscillate near this resonance frequency.

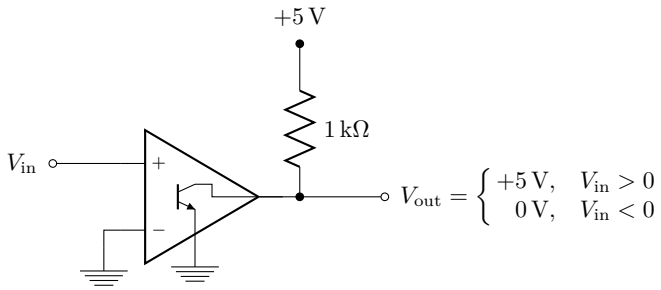
7.9 Comparators

Recall that op-amps are basically high-gain differential amplifiers.



We have mostly concentrated on closed-loop operation (feedback from output to the inverting input), which forces the inputs to have basically the same voltage. In open-loop operation (no feedback), the inputs are not the same, and if they are different by even a small amount ($\sim$mV), the output rails one way or the other to reflect the difference. This open-loop operation is useful in some contexts, and op-amps that are specifically designed for this purpose are called **comparators**.

Specialized comparators (vs. using regular op-amps in the same role) have some advantages. For example, stability is not a concern, because comparators are not generally used with negative feedback. Thus, they need no compensation, and are instead optimized for very high slew rates. In fact, a common configuration for a comparator is the **open-collector output**. The common LM311 comparator, with open-collector output, is shown below, connected as in typical usage.



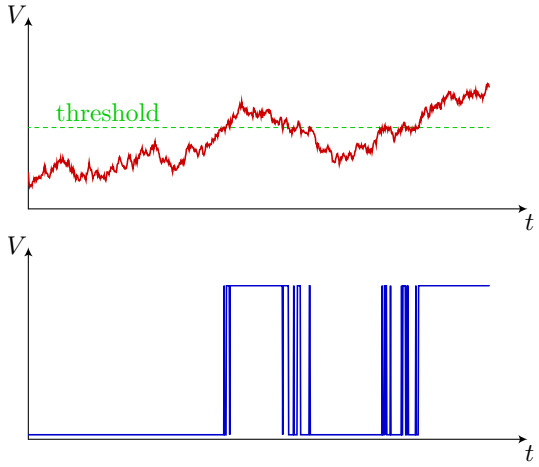
The usual output of the op-amp in the LM311 drives the base of an output transistor, whose collector is connected to the output. If $V_{in} > 0$, then the transistor acts as an open circuit, causing the output to go high to +5 V via the 1 kΩ pull-up resistor. If $V_{in} < 0$, the transistor acts as a short, causing the output to fall to zero.

Comparators are useful in interfacing analog signals to digital circuits, which only recognize two states (`HIGH` voltage and `LOW` voltage). The comparator simply compares the analog signal to some reference voltage, and "tells" the digital circuit whether the analog signal is above or below the reference, but using the correct digital voltages. The states of 0 and +5 V as in the LM311 example above are appropriate for TTL logic, for example. More complex interfaces are certainly possible, and we will return to this later when we discuss analog-to-digital conversion.

Beyond digital interfacing and analog-to-digital conversion, other applications of comparators include oscillators and drivers for alarms or indicators (LEDs, buzzers, beepers) based on an input sensor (e.g., for temperature or water level).

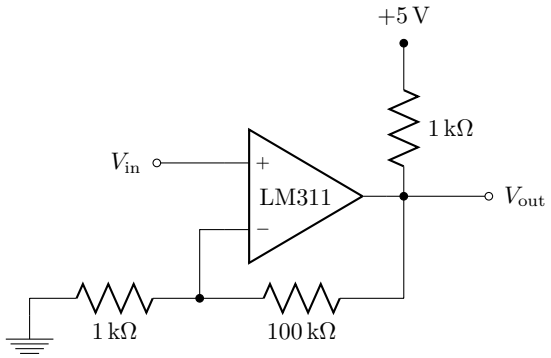
7.9.1 Schmitt Trigger

One problem with comparators arises with noisy input signals. Consider the noisy input voltage below going into a comparator with the reference voltage shown. What we want from the comparator is a signal that reflects when the input signal goes above or below the reference. The corresponding output is shown in the lower graph.



But what we see is that due to the noise, the output signal makes many (spurious) transitions whenever the signal crosses a reference, whereas we would expect a smooth input signal to make only one transition at each crossing.

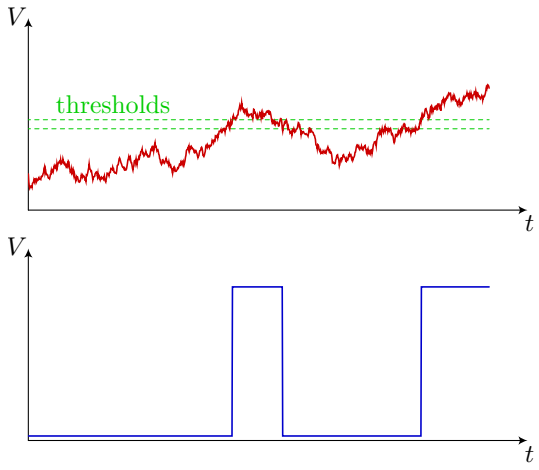
A solution to this is *positive feedback*, which introduces hysteresis. The circuit below, based on the LM311, uses feedback to the noninverting input.



Again, the output swings between 0 V and 5 V, depending on the inputs. Now look at the two cases.

1. If V_{in} is low, then V_{out} is high (+5 V), and the trigger point is about $50\,mV$.
2. If V_{in} is high, then V_{out} is low (0 V), and the trigger point is $0\,mV$.

The trigger point depends on the output, and thus to the input; in other words, V_{in} "repels" the trigger point, and this gives the circuit immunity to noise at the level of about 50 mV or less. The schematic operation of the Schmitt trigger, from introducing the two effective trigger points, is shown below on the same signal. The hysteresis suppresses the spurious transitions. (Note that the output is inverted compared to the discussion of the LM311 circuit, so it compares more closely to the comparator output in the previous graphs.)



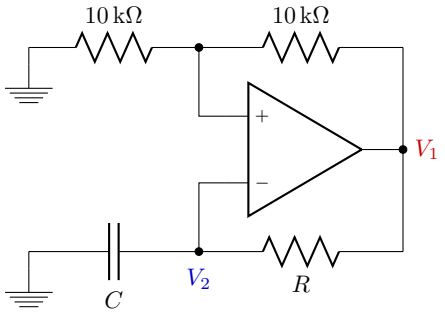
Of course, other nominal trigger levels besides 0 V are possible, by replacing the 1-kΩ resistor with a voltage divider. The Thévenin resistance of the divider acts in place of the 1-kΩ resistor.

7.10 Positive Feedback and Oscillator Circuits

Besides the Schmitt trigger, positive feedback is useful in op-amp oscillators. We will study two examples of positive-feedback oscillators here: a relaxation oscillator and a phase-shift oscillator.

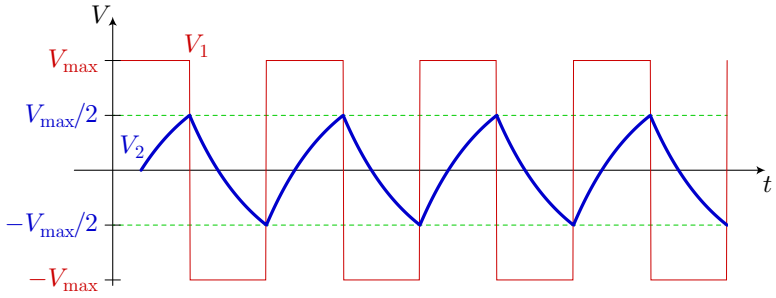
7.10.1 Relaxation Oscillator

One good example of a positive-feedback oscillator is the **relaxation oscillator**, shown below.



Here, the amplifier is standard op-amp, acting as a comparator in open-loop mode. We will assume the output rails are $+V_{max}$ and $-V_{max}$. There is a 50% voltage divider feeding the noninverting input, similar to the Schmitt trigger above. This sets the trigger points of the comparator to $+V_{max}/2$ and $-V_{max}/2$.

Now consider the output of the oscillator at the two points V_1 and V_2, shown below.



If the output V_1 is positive, the RC circuit charges V_2 until the inverting-input voltage exceeds the $V_{\max+}/2$ trigger point, at which point V_1 goes negative, and the charging proceeds in the opposite direction until V_2 reaches $V_{\max-}/2$, and the cycle repeats.

To treat this more quantitatively, the interval between the switching times is the time from RC decay of V_2 from $+V_{\max}/2$ to $-V_{\max}/2$. The process is (RC) exponential decay starting from $+V_{\max}/2$ to $-V_{\max}$, so we are waiting for the decay to 1/3 of the initial voltage, thinking of $-V_{\max}$ as "ground." That is, if Δt is the time interval, then

$$e^{-\Delta t/RC} = \frac{1}{3}, \tag{7.78}$$

so

$$\Delta t = RC \log 3 \approx 1.1\, RC. \tag{7.79}$$

The period T is $2\Delta t$, so we have

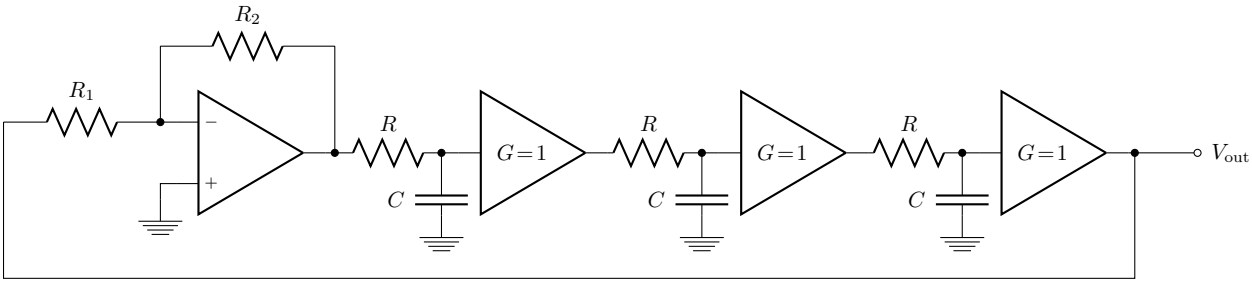
$$T = RC \log 9 \approx 2.2\, RC \tag{7.80}$$

(relaxation-oscillator period)

for the period of the relaxation oscillator. The output can be either a quasi-triangle wave or a square wave, depending on which point serves as the output.

7.10.2 Buffered Phase-Shift Oscillator

Another example of an oscillator is shown below. It produces a sine wave at a frequency determined by the RC low-pass filters. The buffers are op-amps connected as unity-gain followers.



Note that the first op-amp is connected as an inverting amplifier, and the output V_{out} feeds back into the inverting amplifier. There are 3 RC filters in the feedback loop. The oscillation condition is that the phase shift of each RC filter is 60°, so the total RC phase shift is 180°. In combination with the action of the inverting amplifier, this is a total phase of 0°, which means that we have positive feedback.

The correct phase shift only happens at one frequency, which we can find by setting the low-pass-filter phase [Eq. (2.61)]

$$\phi = -\tan^{-1}(\omega RC) \tag{7.81}$$

to $\phi = 60°$. The solution is the angular frequency

$$\omega = \frac{\tan 60°}{RC} = \frac{\sqrt{3}}{RC} \approx \frac{1.732}{RC}. \tag{7.82}$$

This corresponds to a frequency $f = \omega/2\pi$, or

$$f = \frac{\sqrt{3}}{2\pi RC} \approx \frac{0.276}{RC}.$$

(oscillation frequency, phase-shift oscillator) (7.83)

For example, if $R = 10\,\mathrm{k}\Omega$ and $C = 0.01\,\mu\mathrm{F}$, then $f = 2.76\,\mathrm{kHz}$.

Recall that the low-pass amplitude transfer function is [Eq. (2.45)]

$$T(\omega) = \frac{1}{\sqrt{1 + (\omega RC)^2}}. \tag{7.84}$$

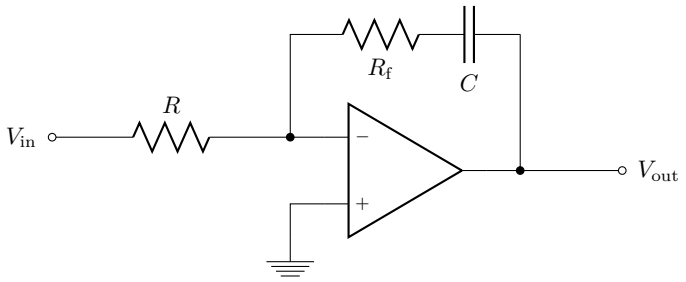
At the oscillation frequency, $\omega RC = \sqrt{3}$, $T(\omega) = 1/2$, so the transfer of 3 RC sections is $1/8$. Thus, to ensure oscillation, we should set $R_2/R_1 = 8$ or a bit higher.

7.11 Circuit Practice

7.11.1 Analog Computers

7.11.1.1 Proportional–Integral Amplifier

As an introduction to the next problem, compute the output voltage in the op-amp circuit below. (It should be proportional to the sum of the integral of the input signal and the input signal itself.) For simplicity, assume $V_{\text{out}}(0) = V_{\text{in}}(0) = 0$.



Solution. Using the inverting-amplifier result,

$$V_{\text{out}} = -\frac{R_f + X_C}{R} V_{\text{in}} = -\frac{R_f}{R} V_{\text{in}} - \frac{i}{\omega RC} V_{\text{in}}. \tag{7.85}$$

Multiplying through by $-i\omega$,

$$-i\omega V_{\text{out}} = i\omega \frac{R_f}{R} V_{\text{in}} - \frac{1}{RC} V_{\text{in}}, \tag{7.86}$$

and then changing to derivatives,

$$\frac{dV_{\text{out}}}{dt} = -\frac{R_f}{R} \frac{dV_{\text{in}}}{dt} - \frac{1}{RC} V_{\text{in}}. \tag{7.87}$$

Integrating,

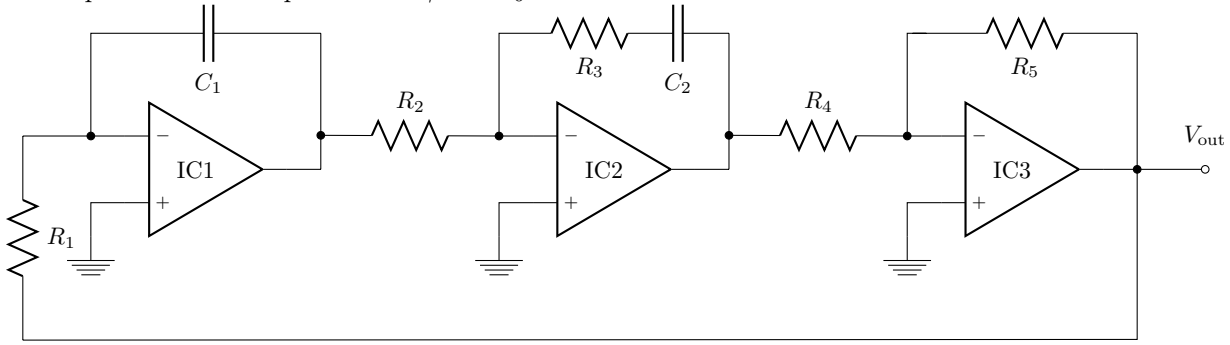
$$V_{\text{out}}(t) = -\frac{R_f}{R} V_{\text{in}}(t) - \frac{1}{RC} \int_0^t V_{\text{in}}(t')\, dt'. \tag{7.88}$$

7.11.1.2 Damped Harmonic Oscillator

The circuit below is an example of an **analog computer**, in this case a computer that solves a differential equation. In particular, show that this circuit solves the damped-harmonic-oscillator equation,

$$\ddot{x} = -\gamma \dot{x} - \omega_0^2 x. \tag{7.89}$$

Give expressions for the parameters γ and ω_0 in terms of the R and C values.



Hint: think of the input to IC1 as $\ddot{x}(t)$, and start integrating from there.

Actually, this circuit only solves for $\ddot{x}(t)$, while $x(t)$ is buried in an inaccessible way in IC2. Can you think of a way to modify the circuit, by replacing IC2 with two other op amps, to make $x(t)$ available?

Note also in particular how R_3 controls the *damping* (γ) for the circuit.

Solution. Suppose the input to the IC1 integrator is $\ddot{x}$. Then the output of IC1 (and the input of IC2) is

$$V_{\text{IC1}} = -\frac{\dot{x}}{R_1 C_1}. \tag{7.90}$$

Now applying the results of the first problem, the output of IC2 is

$$V_{\text{IC2}} = \frac{R_3}{R_1 R_2 C_1}\dot{x} + \frac{x}{R_1 R_2 C_1 C_2}. \tag{7.91}$$

The last op amp just inverts with some gain:

$$V_{\text{IC3}} = V_{\text{out}} = -\frac{R_3 R_5}{R_1 R_2 R_4 C_1}\dot{x} - \frac{R_5}{R_1 R_2 R_4 C_1 C_2}x. \tag{7.92}$$

Then $V_{\text{out}} = \ddot{x}$, so we have the equation

$$\ddot{x} = -\frac{R_3 R_5}{R_1 R_2 R_4 C_1}\dot{x} - \frac{R_5}{R_1 R_2 R_4 C_1 C_2}x. \tag{7.93}$$

This is the harmonic-oscillator equation with

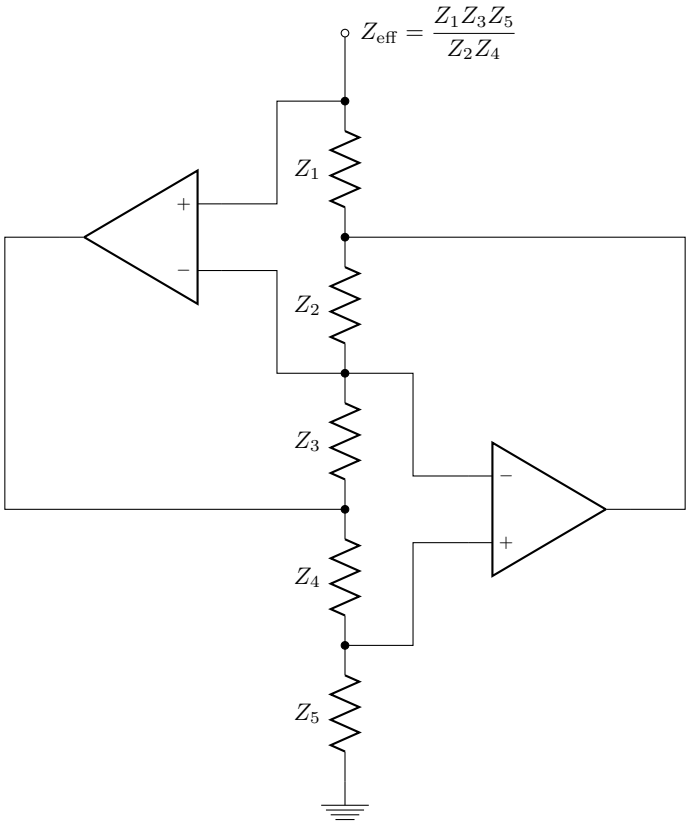
$$\gamma = \frac{R_3 R_5}{R_1 R_2 R_4 C_1}, \qquad \omega_0 = \sqrt{\frac{R_5}{R_1 R_2 R_4 C_1 C_2}}. \tag{7.94}$$

Note that we set V_{out} to $\ddot{x}$, so this solves for the second derivative. To make $x(t)$ available, replace IC2 by two op amps, one an integrator and one an inverting amplifier. Then use IC3 as an inverting summer. The $x(t)$ signal is then available on the output of the inverting amplifier.

7.11.2 Gyrator

The circuit below is an example of a **gyrator**,[1] which presents an effective impedance based on the constituent impedances Z_1–Z_5 (which are not necessarily resistors).

[1] Paul Horowitz and Winfield Hill, *The Art of Electronics*, 2nd ed. (Cambridge, 1989), p. 291 (ISBN: 0521370957).



(a) Show that the effective input impedance is as advertised. Begin by considering a voltage V and a current I at the input terminal, and divide them to obtain the impedance.

Hints: to get you started, here are a few things to simplify and keep things organized. First, consider currents I_1–I_5, flowing through "resistors" Z_1–Z_5, all in the direction of the ground connection.

Second, *assume* that both op amps are operating "normally" with negative feedback (this is *plausible* in this circuit, since the outputs are "closer" in the impedance chain to the negative inputs than to the negative outputs, but strictly speaking this would need to be proven; for simplicity, just assume this is the case).

Third, note that $V = I_5 Z_5$. (Why?)

Finally, use what you know about op amps to relate all the different currents together; you don't need to consider any other currents besides I and I_1–I_5, and obviously you want to eliminate all of them but I.

(b) One of the utilities of this circuit is to realize an effective inductor using only op amps, resistors, and capacitors. This is useful since these components are often better behaved (i.e., closer to ideal) than inductors. Suppose Z_4 is a capacitor in this circuit, with the rest resistors. Show that the result is an effective inductor, and give the effective inductance. What set of (reasonable) components would give you a 1-H inductor? (A pretty big inductor!)

Solution.
(a) We start with

$$V = I_5 Z_5. \tag{7.95}$$

This is because the voltage drop across the other four resistors must be zero, because they are wrapped between the inputs of the op amps.

Now the voltage between the inputs of the right-hand op amp is zero, so

$$I_3 Z_3 = -I_4 Z_4. \tag{7.96}$$

Then using $I_4 = I_5$ (no current into the op-amp input), we find

$$V = I_4 Z_5 = -\frac{Z_3 Z_5}{Z_4} I_3. \tag{7.97}$$

Repeating this argument, we have

$$I_1 Z_1 = -I_2 Z_2 \tag{7.98}$$

and $I_2 = I_3$, so

$$V = -\frac{Z_3 Z_5}{Z_4} I_2 = \frac{Z_1 Z_3 Z_5}{Z_2 Z_4} I_1. \tag{7.99}$$

Finally, $I_1 = I$, so

$$V = \frac{Z_1 Z_3 Z_5}{Z_2 Z_4} I =: I Z_{\text{eff}}, \tag{7.100}$$

which establishes the effective impedance.

(b) With

$$Z_{\text{eff}} = \frac{Z_1 Z_3 Z_5}{Z_2 Z_4} \tag{7.101}$$

and setting $Z_4 = i/\omega C$ and the other impedances to resistances,

$$Z_{\text{eff}} = -i\omega \frac{R_1 R_3 R_5 C}{R_2}. \tag{7.102}$$

Comparing this to

$$X_L = -i\omega L, \tag{7.103}$$

we have an inductance

$$L_{\text{eff}} = \frac{R_1 R_3 R_5 C}{R_2}. \tag{7.104}$$

With $C = 0.01\,\mu\text{F}$, we could pick all resistors to be $10\,\text{k}\Omega$, which would give 1 H of inductance.

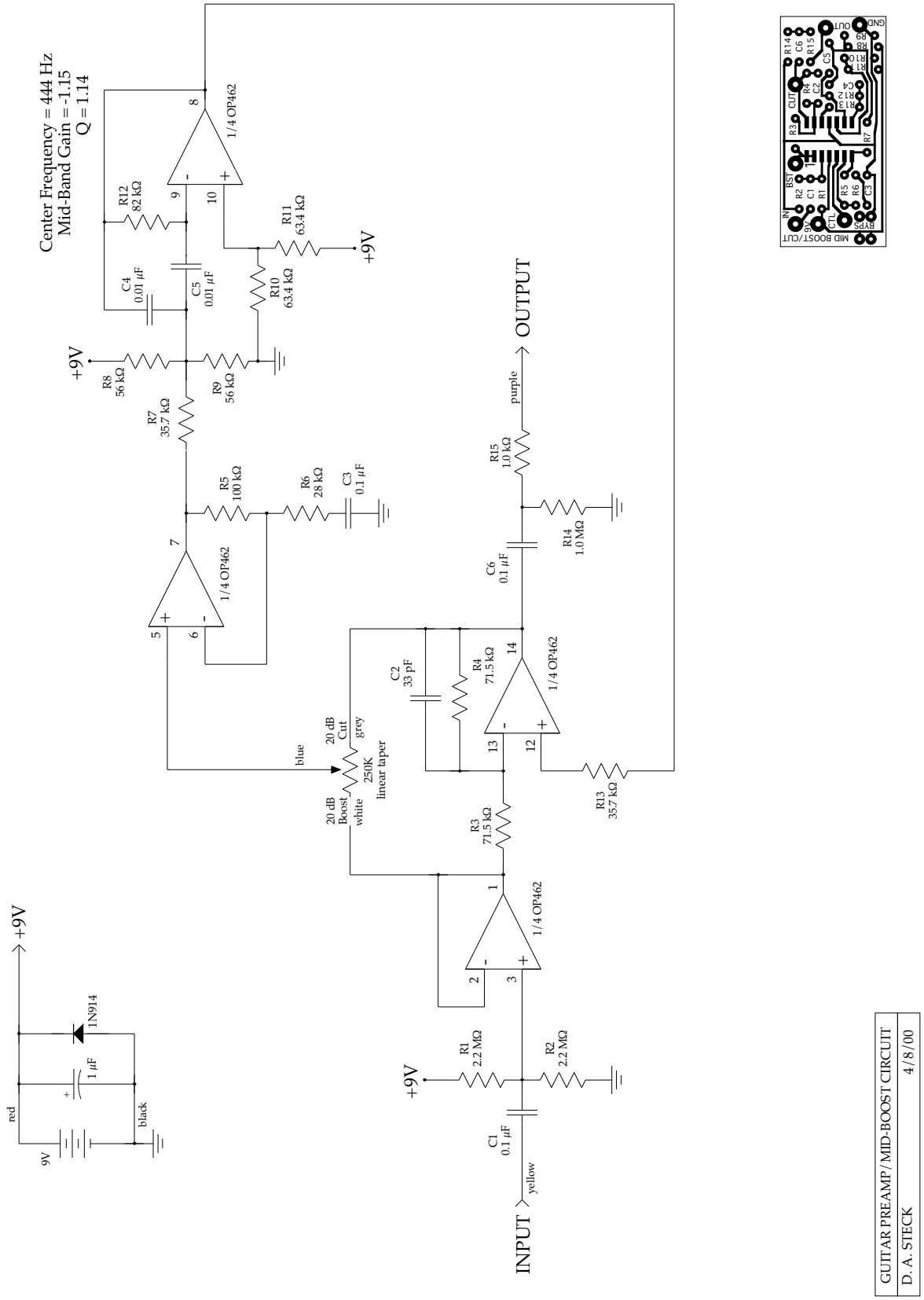
7.11.3 Guitar Preamp with Midrange Boost/Cut

The circuit on the next page is a preamplifier for an electric guitar, powered from a single 9-V battery. It is designed for a Fender Stratocaster, and uses one of the "tone knobs" to control a midrange boost or cut (a midrange boost gives a "fat" sound more like a Les Paul guitar, with "humbucking pickups;" a midrange cut gives a clear, "thin" sound, more like the neck pickup on a Fender Telecaster guitar). The PCB design is also shown, printed at actual size. (Compare this circuit to the preamplifier for the Eric Clapton Signature Stratocaster from Section 4.12.4.)

Try to work through the circuit and understand each of the elements, noting the following:

- Since this circuit is powered by a single 9-V battery, but the signal is bipolar, all circuits must be referenced to the "effective ground" of 4.5 V. For example, the input is ac-coupled and biased at this effective ground.
- There are four op-amps, but all packaged in one chip. Hence the "1/4 OP462," and the IC pin numbers on each op-amp.
- First, convince yourself that the input op-amp (pins 1-3) is an ac-coupled, unity-gain buffer.
- Now, the upper-left op-amp (pins 5-7) functions as a noninverting amplifier. But the capacitor gives a frequency-dependent gain. You should convince yourself that the dc gain is unity, but the ac gain is higher. What is the ac gain? Why do we want unity gain at dc?

- The upper-right-hand op-amp is less obvious, but this is an inverting bandpass filter, as suggested by the center frequency and Q factor. Note that the noninverting input is biased at virtual ground (normally this input would just be grounded), and that there are two R–C pairs at the input (R7 in parallel with R8 and R9, with C4; and R12 with C5) that together give the bandpass action, because they act something like cascaded low-pass and high-pass filters. (See Problem 22.)
- The final op-amp (pins 12-14) combines the filtered signal with the original to give the boost or cut. Capacitor C2 rolls off the gain at high frequencies, where the combination may not be accurate due to different delays at the inputs.
- To control the amount of boost or cut, a potentiometer (with color-coded wires according to the standard Fender convention for the tone knob) interpolates between the buffered input signal and the inverted version produced by the last op-amp.
- The output is ac-coupled and biased to ground.
- Note the diode in the battery/power-supply connection, which protects the op-amp in case the battery is accidentally connected in reverse. A single bypass capacitor stabilizes all op-amps, since they are in a single package. The op-amp operates at low power, so the bypass capacitor value and location are not critical; a larger capacitance is fine since the capacitor does not need to work well at very high frequencies (audio = low frequency in circuits).

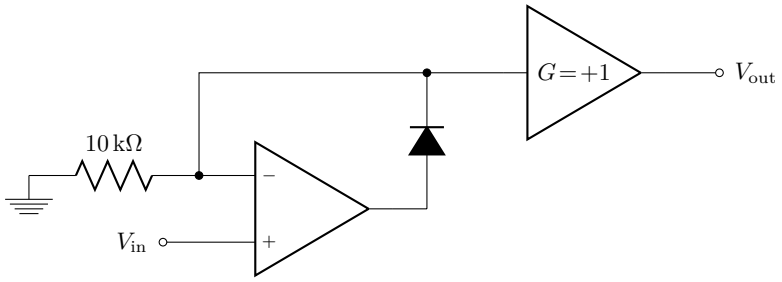

Center Frequency = 444 Hz
Mid-Band Gain = -1.15
Q = 1.14
+9V
INPUT
OUTPUT
yellow
purple
blue
grey
white
red
black
9V
1N914
1 µF
C1 0.1 µF
C2 33 pF
C3 0.1 µF
C4 0.01 µF
C5 0.01 µF
C6 0.1 µF
R1 2.2 MΩ
R2 2.2 MΩ
R3 71.5 kΩ
R4 71.5 kΩ
R5 100 kΩ
R6 28 kΩ
R7 35.7 kΩ
R8 56 kΩ
R9 56 kΩ
R10 63.4 kΩ
R11 63.4 kΩ
R12 82 kΩ
R13 35.7 kΩ
R14 1.0 MΩ
R15 1.0 kΩ
20 dB Boost
20 dB Cut
250K linear taper
1/4 OP462
MID BOOST/CUT
GUITAR PREAMP / MID-BOOST CIRCUIT
D. A. STECK
4/8/00

7.11.4 Active Rectifiers

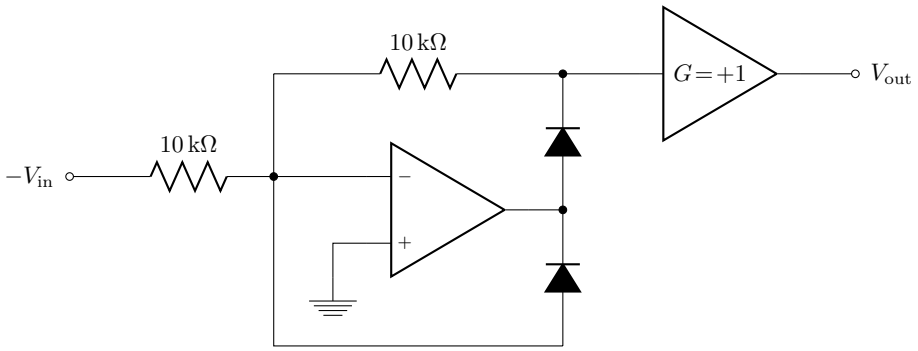
Below are two active rectifiers; that is, ideally, they realize the function

$$V_{\text{out}} = \begin{cases} V_{\text{in}}, & V_{\text{in}} \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7.105}$$

The first one is a "simple" rectifier,



and the second one is the "better" rectifier.[2]



Note that the second circuit takes the *inverted* signal $-V_{\text{in}}$ instead of V_{in}, which could be implemented by another inverting amplifier that is not shown.

The questions are:

1. Why are these active rectifiers? (Note that unlike simple diodes, these circuits really make a transition at $V_{\text{in}} = 0$, rather than at one forward diode-drop above ground.)
2. Why is the "better" circuit better? (*Hint*: it has to do with the slew rate; what is the state of the op-amps when $V_{\text{in}} < 0$?)

Solution. Tracing through the simple rectifier: Note that if $V_{\text{in}} \geq 0$, then the output of the op-amp can maintain the inverting-input voltage $V_{\text{in}-} = V_{\text{in}}$ by keeping its output at one forward diode drop above V_{in}. However, if $V_{\text{in}} < 0$, then the op-amp can't pull $V_{\text{in}-}$ negative through the diode, so the op-amp rails negative. The output is taken from $V_{\text{in}-}$, so the forward diode-drop doesn't matter for calculating V_{out}.

Tracing through the better rectifier: If $-V_{\text{in}} \leq 0$ (i.e., $V_{\text{in}} \geq 0$), then the output of the op-amp can maintain the inverting-input voltage $V_{\text{in}-} = 0$ by pulling its output one diode drop above V_{in} and conducting via the upper diode, and the op-amp acts like an inverting amplifier. Again, the diode drop doesn't matter since the output is buffered at the correct point. For $-V_{\text{in}} \geq 0$ (i.e., $V_{\text{in}} \leq 0$), the lower diode conducts instead, and the upper diode disconnects—this means that $V_{\text{out}} = 0$ because the output buffer sees the virtual ground at the inverting input to the op-amp. The output of the op-amp is one forward diode-drop below ground.

The difference in these circuits is in the crossing through zero, because in the former case, the op-amp output swings from the negative rail to $+0.7\,\text{V}$. In the latter case, the op-amp only swings from $-0.7\,\text{V}$ to $+0.7\,\text{V}$, which reduces the tendency of the op-amp to glitch when fast input signals have zero-crossings.

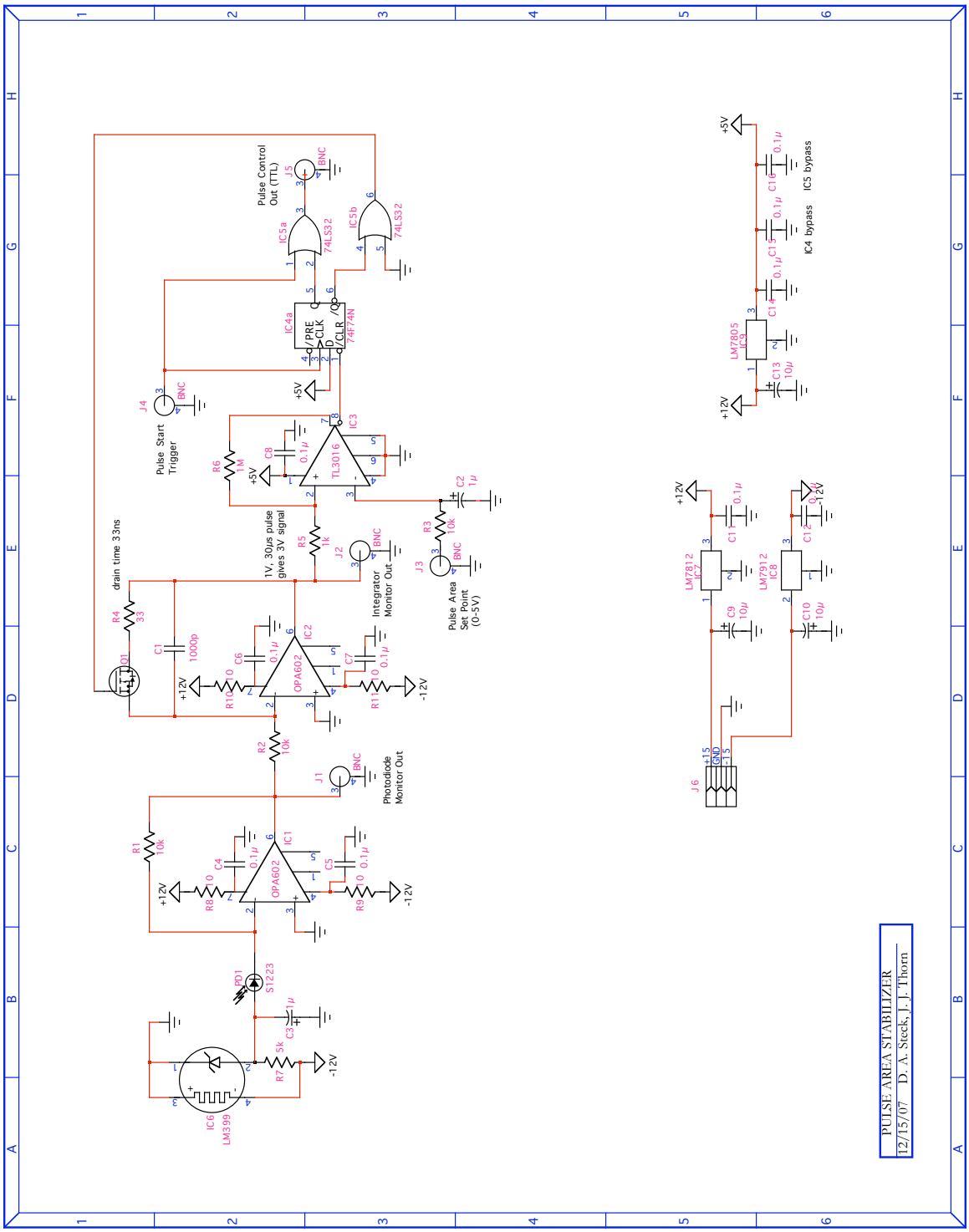
[2]Paul Horowitz and Winfield Hill, *The Art of Electronics*, 2nd ed. (Cambridge, 1989), pp. 187-8 (ISBN: 0521370957).

7.11.5 Pulse-Area Stabilizer

The circuit on the next page is designed for the following purpose. To take photographs with a laser pulse, it it desirable to have the same exposure from each pulse. But the intensity of the laser drifts. Rather than try to stabilize the intensity of the laser, we can compensate for the drift by changing the *duration* of each laser pulse to compensate. By making the **pulse area** or integrated energy of each pulse the same, the photographs have exactly the same exposure, independent of the laser intensity.

Try to trace through the following features in the circuit.

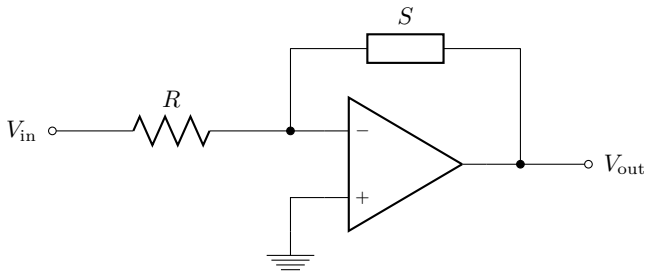
1. The reference IC6 provides a stable voltage to bias the photodiode. What is the voltage at pin 2?
2. You should then convince yourself that the photodiode PD1 is reverse-biased. This helps to improve the speed of the photodiode. The photodiode itself acts as a current source, with current flowing from cathode to anode.
3. What kind of op-amp circuit is IC1? How is the output related to the photodiode signal? (Answer: the op-amp output is positive and proportional to the photocurrent.)
4. IC2 is an integrator, with a MOSFET to reset the integrating capacitor.
5. IC3 is a comparator, connected as a Schmitt trigger. It detects when the integral of the laser pulse intensity reaches a set value from input jack J3. Its output drives digital logic circuitry, to which we will return later after we have studied digital electronics.


PULSE AREA STABILIZER
12/15/07 D. A. Steck, J. J. Thorn
Pulse Control Out (TTL)
Pulse Start Trigger
drain time 33ns
1V, 30μs pulse gives 3V signal
Integrator Monitor Out
Pulse Area Set Point (0-5V)
Photodiode Monitor Out
IC4 bypass
IC5 bypass
OPA602
TL3016
74LS32
74F74N
LM7812
LM7912
LM7805
LM399
S1223

7.12 Exercises

Problem 7.1

Consider the following op-amp circuit, which contains one resistor (of resistance R), and one schmesistor.



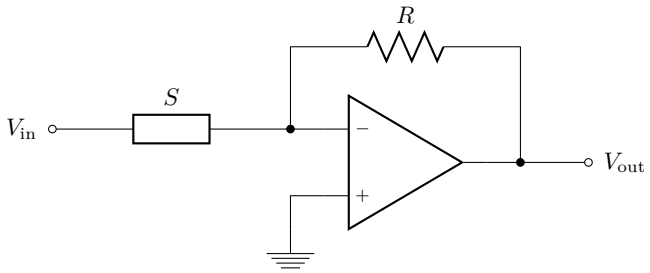
Recall (Problem 12) that a "schmesistor" is a device that obeys "Schmohm's law,"

$$V = I^2 S, \tag{7.106}$$

where S is the "schmesistance." Derive an expression for V_{out} in terms of V_{in}, R, and S.

Problem 7.2

Consider the following op-amp circuit, which contains one resistor (of resistance R), and one "schmesistor."



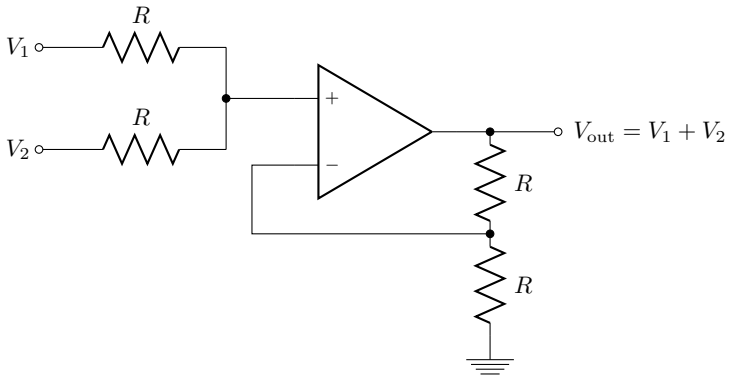
Recall that a "schmesistor" is a device that obeys "Shmohm's law,"

$$V = I^2 S, \tag{7.107}$$

where S is the "schmesistance." Derive an expression for V_{out} in terms of V_{in}, R, and S.

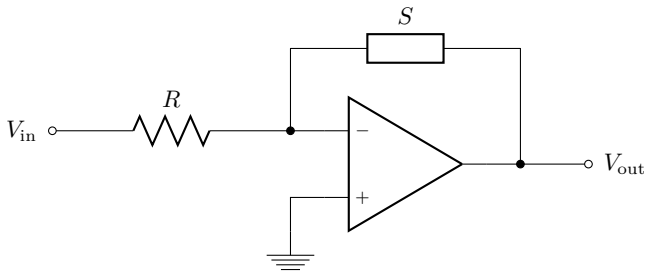
Problem 7.3

Show that $V_{\text{out}} = V_1 + V_2$ in the circuit below.



Problem 7.4

(a) Consider the following op-amp circuit, which contains one resistor (of resistance R), and one schmapacitor (see Problem 9) of schmapacitance S.



Derive an expression for $V_{\text{out}}(t)$ in terms of an arbitrary, time-dependent input $V_{\text{in}}(t)$, R, and S. Assume an ideal op-amp.

Recall that the schmapacitor is defined by the relation

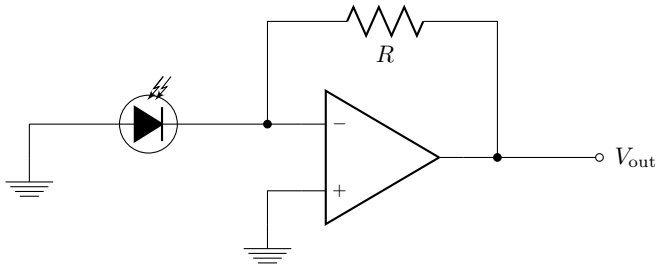
$$I = S\frac{d^3V}{dt^3}, \tag{7.108}$$

where I is the schmapacitor current and V is the voltage drop across the schmapacitor.

(b) Of course, schmapacitors aren't real. Describe *briefly* and *qualitatively* how you could build an equivalent circuit using real-world components.

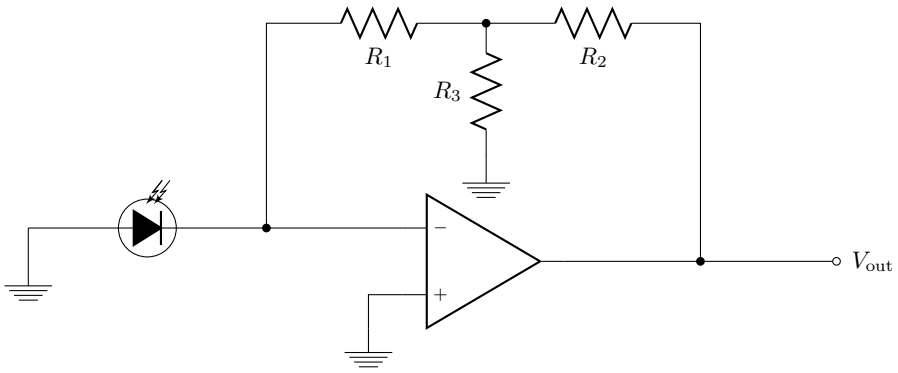
Problem 7.5

(a) A photodiode produces a *backwards* current (i.e., current flows from cathode to anode) when detecting light. (Think of this as the opposite of an LED, where a *forward* current causes light to be *emitted*. In fact, an LED can work as a photodiode, though not a particularly great one.) Consider the photodiode-amplifier (op-amp) circuit below, which acts as a **transimpedance amplifier** (current input, voltage output). Write down an expression for the output voltage in terms of the photodiode current and R. Is the output voltage positive or negative when you shine light on the photodiode?



(b) The Hamamatsu S1223 is a standard, medium-area (2.4 mm×2.8 mm), general-purpose, silicon PIN photodiode. The sensitivity is specified at 0.52 A/W. Assuming the photodiode collects all the power from a steady, 1-μW laser beam, and the resistor is $R = 10\,\text{k}\Omega$, what is the output voltage?

(c) For a very sensitive circuit (i.e., to register small input powers, of the order of nW), it may be necessary to use a very large resistor. But also recall that op-amps aren't always happy with feedback resistances much over 1 MΩ, and very large resistors may be difficult to source. There is a nice trick to get around this, however. Consider the modified transimpedance amplifier below, with a "T network" in the feedback loop.



What is the effective feedback resistance R of this circuit (i.e., the resistance that makes this equivalent to the original circuit)? Find a combination of 100 kΩ or smaller resistors that gives an effective R of 1 GΩ.

Problem 7.6

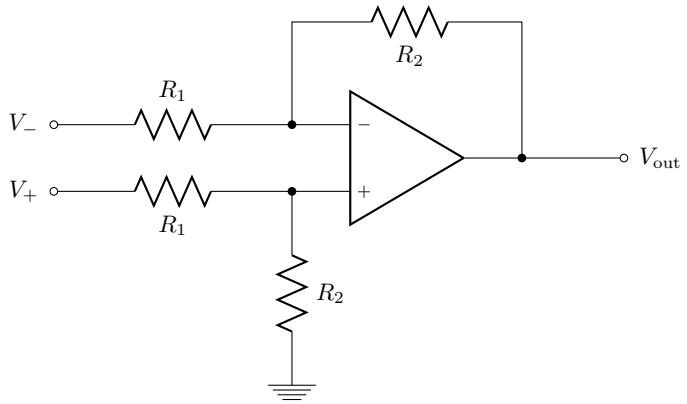
Design a current source (constant-current regulator), according to the following specs:

1. The only allowed parts are: one (ideal) op-amp, one resistor, and the load.
2. The load current is controlled by an input voltage, with a 1-V input change corresponding to a 1-mA change in output current.
3. The load need not be ground-referenced (i.e., the load need not have any direct connection to ground, or to any particular voltage).

Hint: think about how the standard inverting op-amp circuit works.

Problem 7.7

Consider the (op-amp) differential-amplifier circuit shown below. Recall that for this to behave as a good differential amplifier (i.e., for perfect common-mode rejection), the two R_1–R_2 resistor pairs must be matched perfectly (in terms of ratio), assuming ideal op-amp behavior.



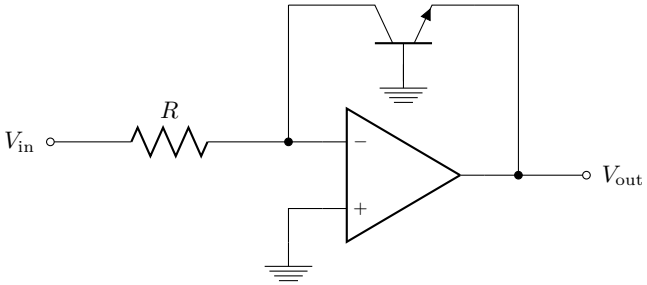
Of course, real resistors aren't perfectly matched. As a model for this, suppose that the feedback resistor has a resistance $R_2 + \delta R$, where δR is a small perturbation to this resistance.

(a) Rederive an expression for V_{out} in terms of the input voltages and resistances. Keep only first-order terms in δR.

(b) Write down an expression for the CMRR. Consider a unity-gain amplifier with $R_1 = R_2 = 10\,\text{k}\Omega$. Give a numerical estimate (in dB) for the expected CMRR if you use 1% resistors. Repeat for 0.01% resistors.

Problem 7.8

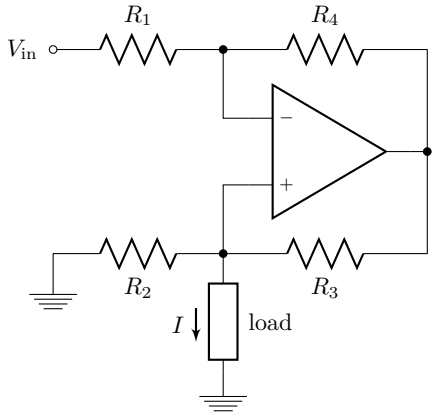
Show that the op-amp circuit below behaves approximately as a **logarithmic amplifier**. Under what conditions does this circuit really function logarithmically?



Hint: the function of the transistor, as in the inverting amplifier, is to convert a current into a voltage, so use an appropriate relation to describe the transistor.

Problem 7.9

For the circuit below, the **Howland current source**, show, provided $R_4/R_1 = R_3/R_2$, that $I = -V_{\text{in}}/R_2$.

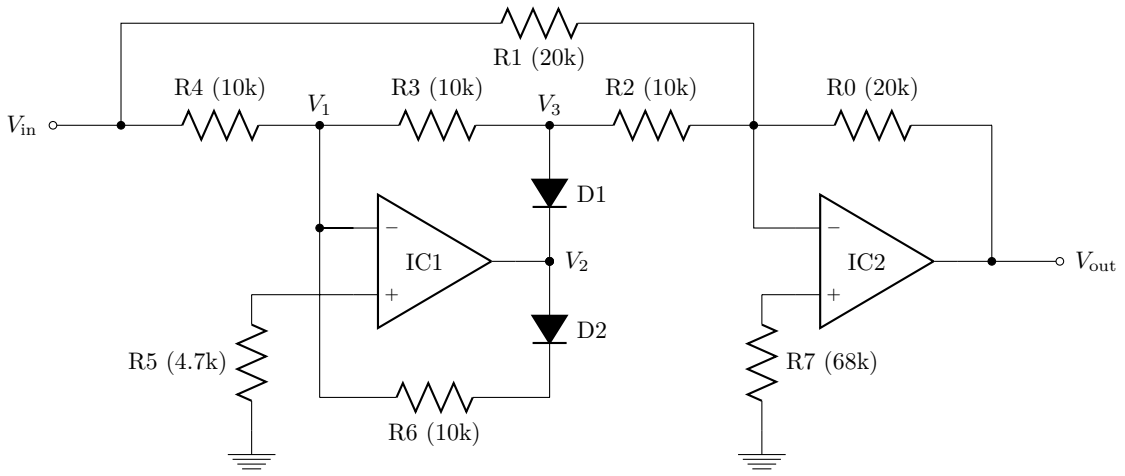


Assume an ideal op-amp.

Problem 7.10

Consider the circuit below.[3] Show that for this circuit, $V_{\text{out}} = |V_{\text{in}}|$.

[3]Miles A. Smither, "Improved absolute-value circuit," in Bill Furlow, Ed., *Circuit Design Idea Handbook* (Cahners Books, 1974), p. 13 (ISBN: 0843602058).



This circuit looks a little complicated, so here is some guidance. Note that you should be able to do this with very little math, provided you break the circuit down into manageable parts that you have already learned about.

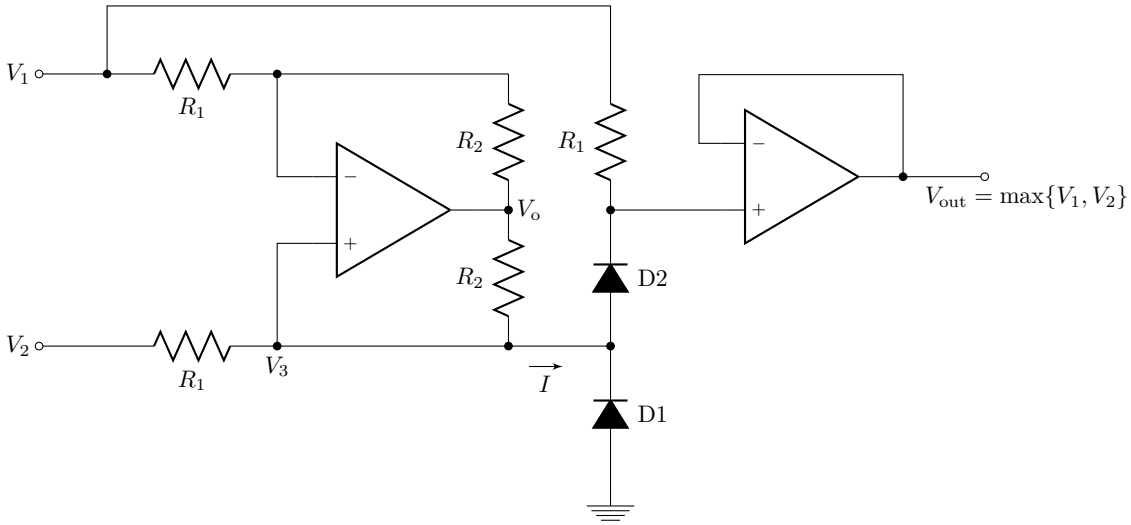
(a) The IC2 op amp is connected in one of the basic op-amp circuits; what is it?

(b) Now the tricky part is understanding the diode network in the IC1 circuit. Begin by treating D1 and D2 as ideal diodes (i.e., no forward voltage drop). Now you should see that only one diode conducts at a time, and in either case IC1 is connected as one of the basic op-amp circuits (which one?). Work out the voltages V_1, V_2, and V_3, and handle separately the cases where the output of IC1 is positive or negative.

(c) Finally, consider the original circuit with *real* diodes, and argue that the forward voltage drops don't matter.

Problem 7.11

The purpose of this problem is to show that, in the circuit below, that $V_{out} = \max\{V_1, V_2\}$ (i.e., it selects the larger of the input voltages). The diodes here are *real* diodes (i.e., you must account properly for any voltage drops across the diodes), and there are no restrictions on the signs of the input voltages.



(a) Begin by showing that the current I is given by

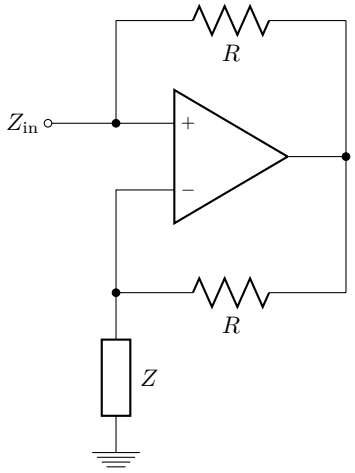
$$I = \frac{V_2 - V_1}{R_1}. \tag{7.109}$$

(The voltage labels V_3 and V_o are there for a reason!)

(b) Then use the current I to show that $V_{\text{out}} = \max\{V_1, V_2\}$.

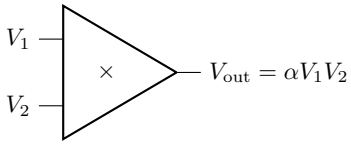
Problem 7.12

The circuit below is a **negative-impedance converter**, composed of an op-amp (you may assume the golden rules apply here), two identical resistors of resistance R, and a generic circuit element of impedance Z (could be a resistor, capacitor, etc.). What is the impedance Z_{in} at the input terminal?



Problem 7.13

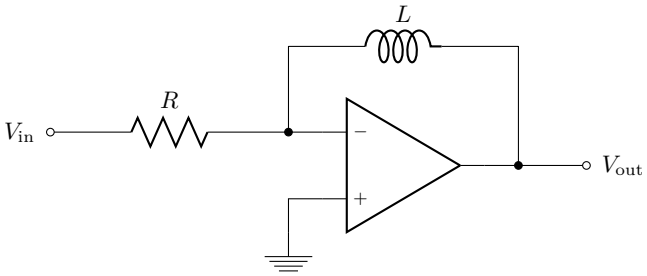
The amplifier below is a multiplying amplifier, with the output voltage proportional to the product of the input voltages.



Using this multiplier and an op-amp, design a circuit that behaves as a **square-root amplifier** (i.e., with $V_{\text{out}} \propto \sqrt{V_{\text{in}}}$). Show that your circuit behaves as advertised.

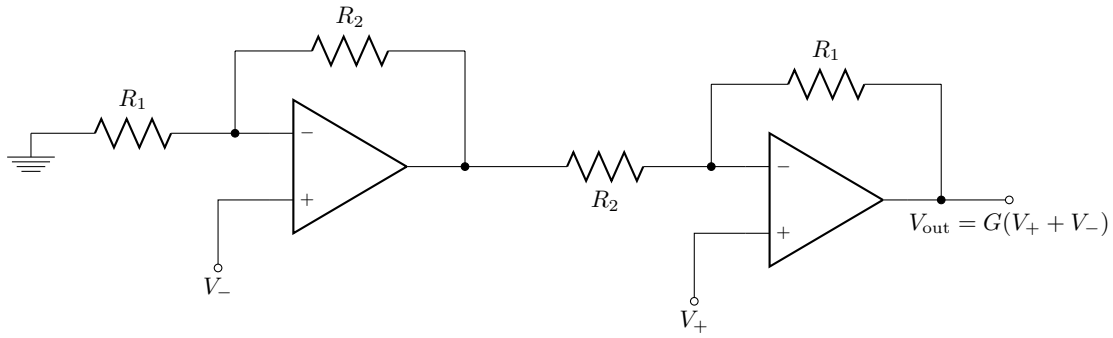
Problem 7.14

Consider the circuit below. Design an equivalent circuit using an op-amp, the **same** resistor R, and a **capacitor**, and give the value C of the capacitance in terms of R and L.



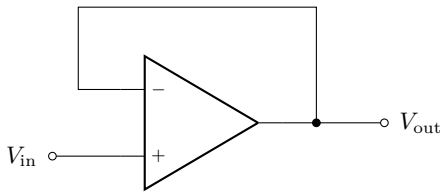
Problem 7.15

Show that the circuit below behaves as an instrumentation amplifier with high input impedance and gain $G = 1 + R_1/R_2$.



Problem 7.16

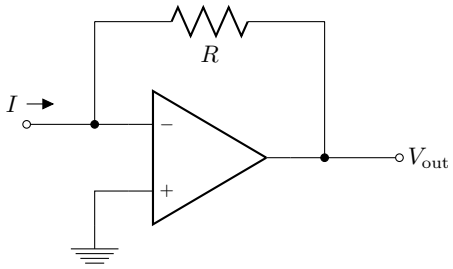
An op-amp unity-gain buffer is shown below.



Compute the closed-loop gain of this circuit, assuming a finite, open-loop gain A.

Problem 7.17

Consider the transimpedance amplifier below. The op amp has **finite** open-loop gain A.



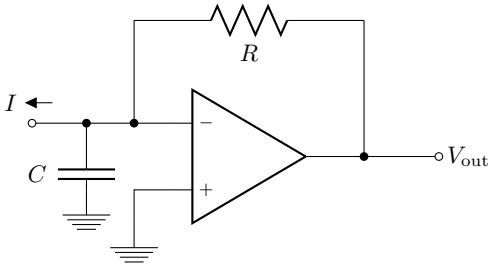
(a) Derive an expression for V_{out} in terms of the input current I. Also take the limit as $A \longrightarrow \infty$.

(b) Derive an expression for the input impedance of the circuit. Ignore any intrinsic input impedance R_{i} of the op-amp inputs.

(c) Derive an expression for the output impedance of the circuit.

Problem 7.18

Consider the transimpedance amplifier below with input capacitance C. This circuit acts as a model for instability in a photodiode ampifier, where recall that a photodiode acts as a current source, and C models the photodiode capacitance.



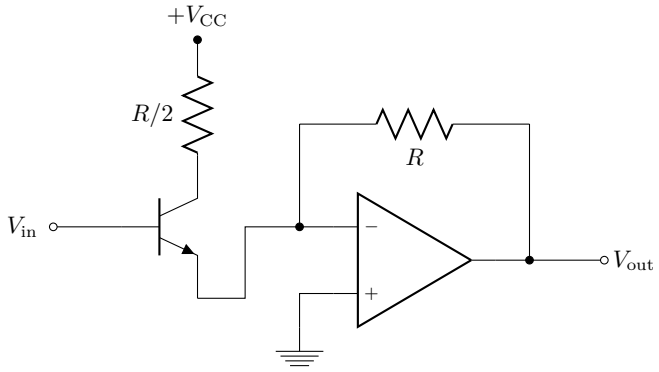
(a) Derive an expression for V_{out}, assuming a **finite** open-loop op-amp gain A. (a) Derive an expression for the frequency-dependent transimpedance $Z(\omega)$ of the amplifier (i.e., such that $V_{\text{out}} = IZ(\omega)$ for an input current of frequency ω). Also take the limit as $A \longrightarrow \infty$.

(b) Show that, if the open-loop gain falls off like a low-pass filter, $A(\omega) = A_0/(1 - i\omega/\omega_0)$, that the magnitude of $Z(\omega)$ is peaked at some frequency $\omega > 0$ if A_0 is sufficiently large.

(c) Find the peak frequency.

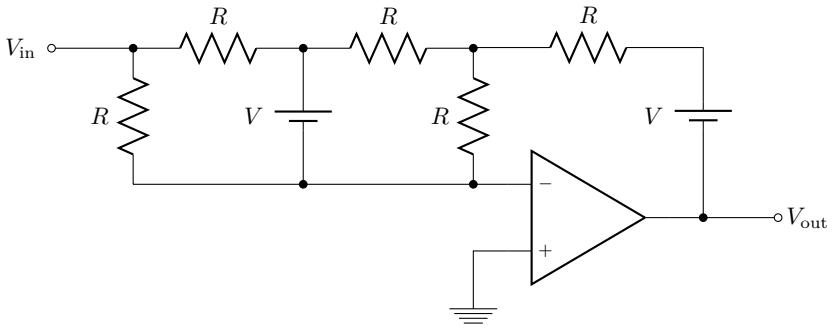
Problem 7.19

Derive an expression for V_{out} in the circuit below, in the regime where the output is not railed (assume that R is of the order of $10\,\text{k}\Omega$). Also, you may assume $V_{\text{in}} \geq 0$ (but assume V_{in} is not so large that it damages the transistor) and an ideal op amp. Finally, the op amp is powered from $\pm V_{\text{CC}}$. Be clear about any assumptions you make.



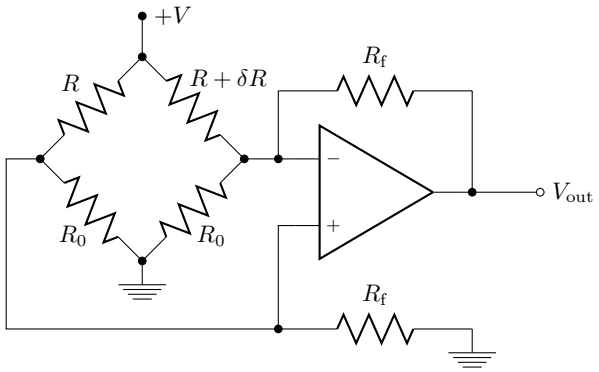
Problem 7.20

In the circuit below, derive an expression for V_{out}, assuming an ideal op amp.



Problem 7.21

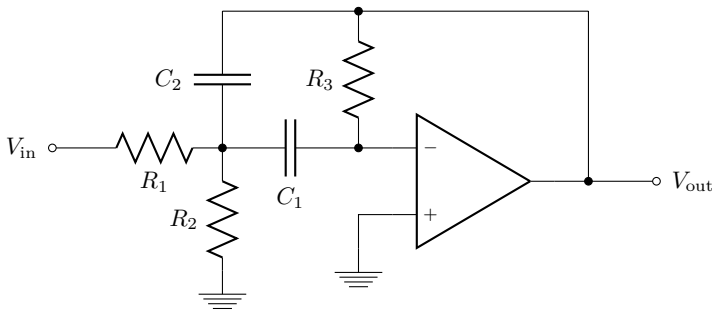
In the circuit below, the diamond-shaped arrangement of resistors is called a **Wheatstone bridge**, and can act as a sensitive measure of the mismatch of two resistors (here R and $R+\delta R$). This is useful, for example, in sensing the value of a thermistor (semiconductor resistor whose resistance varies with temperature).



For this circuit, derive an expression for V_{out}. To keep the algebra under control, give the answer only to lowest order in the small perturbation δR [i.e., give the lowest-order term in the Taylor expansion about $\delta R = 0$; you will need to use $(a + \delta a)^{-1} \approx a^{-1} - \delta a/a^2$ for $\delta a \ll a$].

Problem 7.22

Consider the following circuit, an active, inverting, band-pass filter. (This is the bandpass filter from the guitar preamplifier in Section 7.11.3.)



(a) Derive an expression for the gain function $\tilde{G}(\omega)$, and the amplitude gain function $G(\omega)$. Note that these are the same as the transfer function and amplitude transfer function, respectively, in the passive-filter case.

Note: an algebra program like *Mathematica* will make this problem considerably simpler. Contact me if you need help with learning how to use it for algebra or making plots.

(b) Derive asymptotic expressions for $G(\omega)$ for large and small frequencies. From these expressions, argue that this is a band-pass filter.

(c) Find an expression for the center frequency, where $G(\omega)$ is maximum. Give a numerical value for this frequency (give the frequency in Hz, not rad/s), for the component values $R_1 = 35.7\,\text{k}\Omega$, $R_2 = 28\,\text{k}\Omega$, $R_3 = 82\,\text{k}\Omega$, $C_1 = C_2 = 0.01\,\mu\text{F}$.

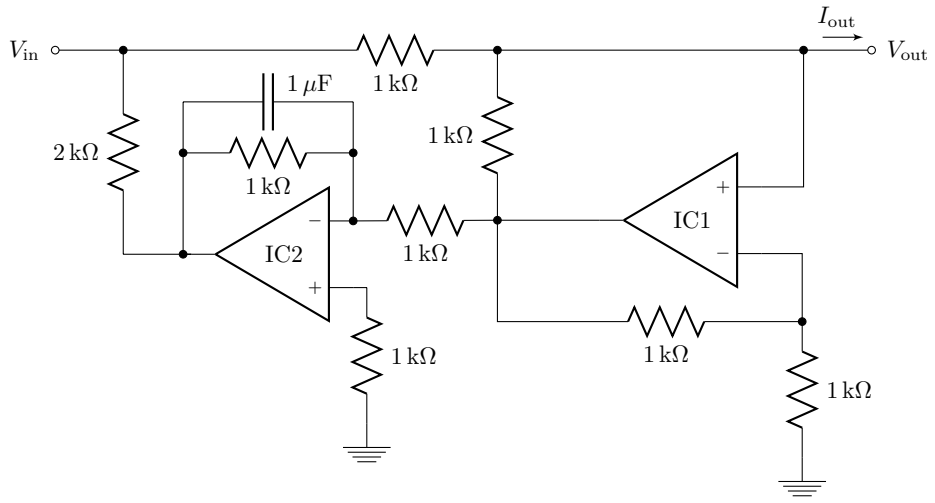
(d) Find an expression for the peak gain, i.e., the value of $G(\omega)$ at the frequency you derived in part (c). Again, give a numerical value for the same component values.

(e) Make a log-log plot of $G(\omega)$, over a reasonable range of frequencies, for the component values above. (Keep in mind this is intended as an audio band-pass filter.)

Problem 7.23

Consider the circuit below, which is intended as a voltage-controlled current source.[4] Assume the "output voltage" V_{out} is held to a fixed voltage by an external source.

[4] This is the modulation input stage of the current-controller circuit in K. G. Libbrecht and J. L. Hall, "A low-noise high-speed diode laser current controller," *Reviews of Scientific Instruments* **64**, 2133 (1993).



(a) First, start by identifying the basic op-amp circuits for IC1 and IC2 (i.e., these are two standard op-amp circuits, connected via a network that includes three resistors). For IC2, first think about the circuit at dc (i.e., ignoring the capacitor).

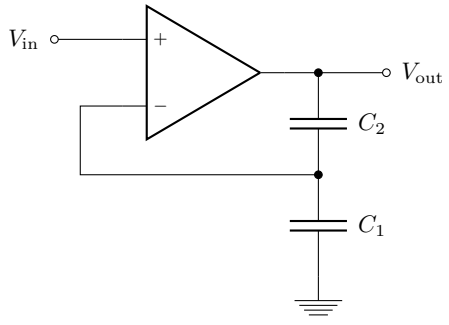
(b) Show that $I_{\text{out}} = V_{\text{in}}/(1\,\text{k}\Omega)$, and is independent of V_{out}, for dc inputs. (You may find it useful to label the output voltages of IC1 and IC2 as V_1 and V_2, respectively.)

(c) Show that if V_{in} is disconnected (i.e., not held at any particular voltage), that $I_{\text{out}} = 0$, independent of V_{out} (for dc inputs).

(d) The function of the capacitor is as follows. The above current regulation requires IC2 to generate a signal to cancel any currents drawn by IC1. However, due to propagation delays through IC2, the cancellation may not be accurate at high frequencies, and in the worst case, the circuit may even become unstable. Thus, the capacitor is there to roll off the gain of IC2, protecting against these effects. However, the cancellation no longer works, so redo (c), calculating I_{out} in terms of V_{out} at high frequencies (i.e., assume that I_{out} and V_{out} are the amplitudes of high-frequency, oscillating signals).

Problem 7.24

Consider the circuit below.

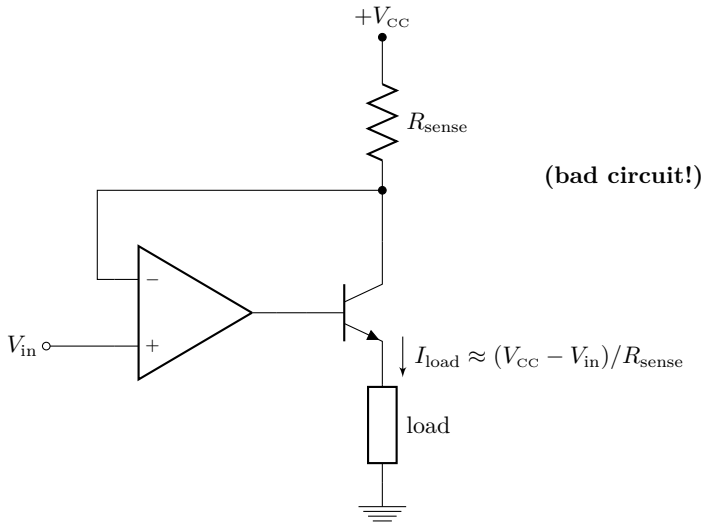


(a) Assuming an *ideal* op-amp, what kind of amplifier is this? Compute V_{out} in terms of V_{in}.

(b) For a *real* op-amp, what would the circuit do [i.e., why wouldn't it work as in (a)], and what *specifically* is it about the op-amp that causes the circuit to misbehave?

Problem 7.25

Consider this op-amp/BJT current-source circuit.



(a) *Assuming* the op-amp golden rules apply to this circuit, show that the expression for the current is approximately correct, and give a *better* expression that includes the transistor β.

(b) As drawn, the circuit does not work; rather, the op-amp just rails without regulating the current. Explain why.

(c) Propose two fixes for this circuit; at least one of these fixes should not involve any component changes.

Problem 7.26

In this problem, you will analyze a schematic for a precision current source that powers a laser diode (see p. 6 in the referenced pdf file).[5] You will probably need to look up data sheets for various components in analyzing this circuit.

(a) Suppose $R_{limit} = 10\,k\Omega$. What is the maximum possible (dc) voltage at pin 3 of the AD820? Assume the 2N7000 is an open circuit. This is the control voltage that sets the current through the laser diode. Note that the 20-kΩ wiper of the (course adjust) pot is bypassed by a capacitor to ensure the control voltage has little high-frequency noise.

(b) Now look at the relay on the left-hand side of the schematic. This is the "laser enable" part of the circuit, which either passes or overrides (i.e., sets to zero) the control voltage. Note the two *momentary* switches (i.e., they are push-buttons, only connected while you are actually pushing them but the circuit "remembers" the last one you pushed). Explain how this section of the circuit works to enable and disable the laser. Include the operation of the status LED (why doesn't the LED burn out if it is powered by 15 V?), and explain why this circuit has a "soft-start" feature.

(c) Now analyze the AD820, which regulates the diode current according to the control voltage at pin 3. For the purposes of this analysis, ignore C0, C1, and R1 (i.e., replace them by open connections) as well as R0, R2, and C2 (i.e., replace them by short circuits). These are needed for stability, but I haven't discussed much about this yet. Thus, the feedback loop consists of the BUF634's, resistor SR10, and the INA128 (assume the gain of the INA128 is 10). What is the purpose of the BUF634's, why are there 2 of them, and what is with the 10 Ω resistors? What is the maximum current through the laser diode given your answer in (a)? *Explain.*

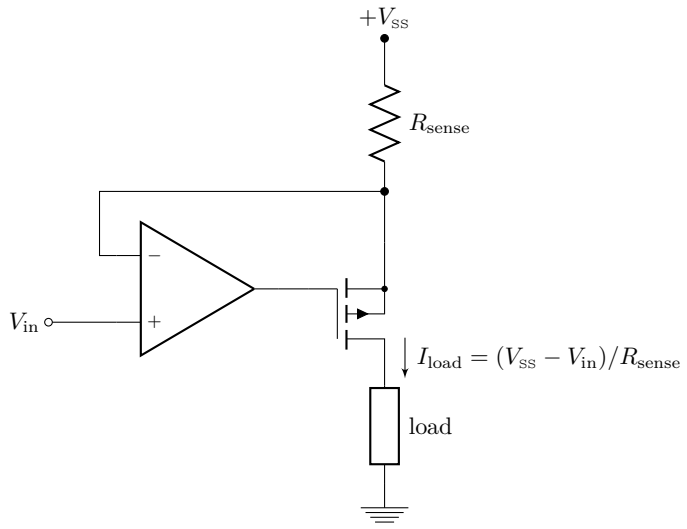
(d) Look at the power-supply connections of the various amplifiers. Explain the differences in the bypass circuits, and *why* the chips are bypassed in different ways (i.e., why not bypass them all in the same way?).

[5] Designed by Todd Meyrath, `http://george.ph.utexas.edu/~meyrath/informal/laser%20diode.pdf`

(e) Now go back to the AD820, and consider it separately from the other ICs, but this time *with* R0-R2, and C0-C2. Also assume pin 3 is grounded. What is the gain at low frequencies? High frequencies?

Problem 7.27

Consider the op-amp current source shown below.



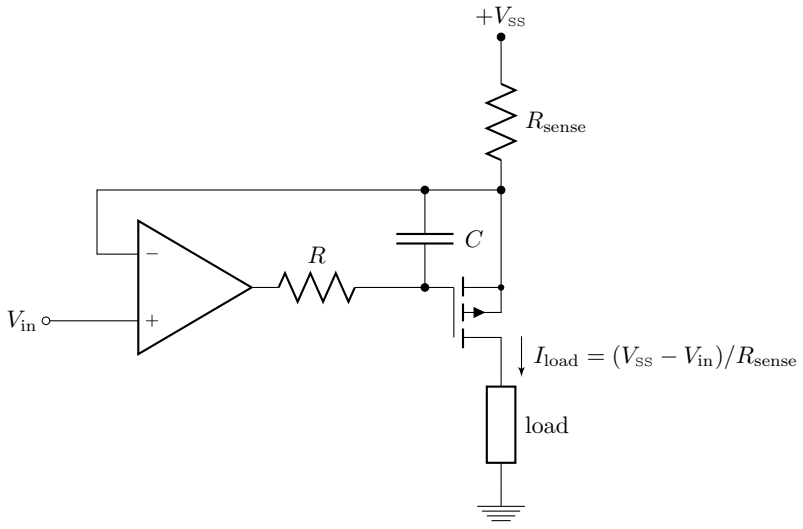
(a) Show that

$$I_{load} = \frac{V_{SS} - V_{in}}{R_{sense}}, \tag{7.110}$$

as advertised.

(b) Note that the MOSFET is a p-channel device. Explain why an n-channel device doesn't work here (consider how the output changes as I_{load} increases).

(c) To consider stability issues in this circuit, consider the modified circuit below, with output resistor R (modeling the intrinsic output resistance of the op-amp), and capacitor C (modeling the gate-source capacitance of the FET; for simplicity we are excluding the gate-drain capacitance, but this could be modeled in the following analysis with another capacitor).



Consider a small ac signal v_{in} at frequency ω at the input (and some dc bias V_0, so $V_{\text{in}} = V_0 + v_{\text{in}}$). Treat the FET via its transconductance g_{m} (i.e., $i_{\text{SD}} = -g_{\text{m}} v_{\text{GS}}$) and the op-amp via the finite-gain formula (7.35)

$$V_{\text{out}} = \tilde{A}(V_{\text{in}+} - V_{\text{in}-}), \tag{7.111}$$

with response

$$\tilde{A}(\omega) = \frac{A_0}{1 - i\omega/\omega_{\text{c}}}, \tag{7.112}$$

where ω_{c} is the cutoff frequency. Show that the ac load response i_{load} to v_{in} can be written

$$i_{\text{load}} = \frac{-g_{\text{m}} A_0}{(1 - i\omega RC)(1 - i\omega/\omega_{\text{c}}) + g_{\text{m}}(1 + A_0 - i\omega/\omega_{\text{c}})R_{\text{sense}}}\, v_{\text{in}}. \tag{7.113}$$

(d) Note that the expression above can change sign, which we will see can lead to instability. Intuitively, the op-amp output is inductive, forming a destabilizing resonance with the capacitive gate input of the FET. Show that the circuit exhibits resonance behavior by finding the extreme value of $|i_{\text{load}}|$, as well as the frequency at which the extremum occurs. Discuss the dependence of the peak frequency on the FET parameters g_{m} and C (simplify your analysis by working in the regime $A_0 \gg 1$ and $\omega_{\text{c}} \ll 1/RC$). Also simplify your notation by defining the frequency ω_0 by

$$\omega_0^2 := \frac{\omega_{\text{c}}}{RC}. \tag{7.114}$$

(e) To more precisely consider stability problems in this circuit, we must consider the *voltage* feedback. To do this we should consider the voltage v_{S} at the inverting input, given by

$$v_{\text{S}} = -i_{\text{load}} R_{\text{sense}} =: G(\omega)\, v_{\text{in}}, \tag{7.115}$$

to be the circuit "output," which then defines the voltage gain G of the circuit. Since this is fed back to the inverting input, G should be positive and preferably large. The problem comes when $\text{Re}[G]$ is *negative*, because it means the feedback is *positive*. In particular, the circuit becomes unstable when $\text{Re}[G(\omega)] < -1$, because perturbations at any frequency satisfying this condition will grow. Find the extreme value of $\text{Re}[G]$ and show that $\text{Re}[G] < -1$ when

$$\frac{A_0 \omega_{\text{c}} RC}{\sqrt{g_{\text{m}} R_{\text{sense}}}(\sqrt{4A_0 \omega_{\text{c}} RC} + \sqrt{g_{\text{m}} R_{\text{sense}}})} > 1, \tag{7.116}$$

again in the limits $A_0 \gg 1$ and $\omega_{\text{c}} \ll 1/RC$. Using typical parameters for a high-speed op-amp and a power MOSFET ($\omega_{\text{c}}/2\pi = 100\,\text{Hz}$, $R = 50\,\Omega$, $R_{\text{sense}} = 50\,\Omega$, $g_{\text{m}} = 1\,\text{℧}$, $A_0 = 5 \times 10^6$) solve the equation to find the range of gate–source capacitance that destabilizes the circuit. Since typical gate capacitances are on the order of 100 pF, you should find that the circuit is unstable (i.e., it oscillates). More components in the feedback loop are necessary to stabilize it.

Again, to keep your calculations organized, you should use the following definitions to simplify your calculation:

$$\begin{aligned} G_0 &:= g_{\text{m}} R_{\text{sense}} A_0 \omega_0^2 \\ \omega_1 &:= \omega_0 \sqrt{1 + g_{\text{m}} R_{\text{sense}}(1 + A_0)} \\ \gamma_1 &:= \frac{\omega_{\text{c}}^2 + (1 + g_{\text{m}} R_{\text{sense}})\,\omega_0^2}{\omega_{\text{c}}}. \end{aligned} \tag{7.117}$$

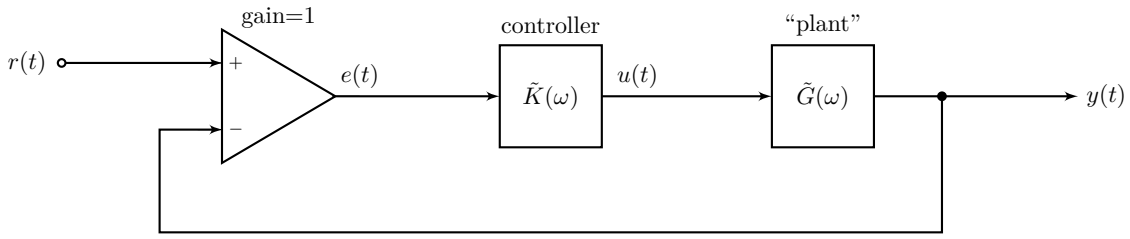
Chapter 8

PID Control

A major theme in our study of op-amps is that **feedback**, and *negative* feedback in particular, is a useful tool for improving the behavior of amplifiers. It is also useful in the realization of circuits that would otherwise be complex or difficult to implement (the logarithmic amplifier is a good example; see Problem 8). Of course, feedback is similarly useful beyond op-amp circuits, and we will consider feedback control more generally as a tool for maintaining systems in a desired "target" state. As it turns out, op-amp circuits are useful in realizing one of the popular, general-purpose control methods, **PID control**, which we will define shortly.[1]

8.1 Basics of Linear Control

Schematically, we can represent a feedback-control system as in the diagram below.



There are several important elements that interact here.

- The **plant** is the system to be controlled. (That's plant as in "chemical plant," not a shrub.) We assume the plant to have an input and an output. The output is some scalar quantity that we want to control, such as temperature, position, voltage, frequency, speed, etc. The input is some other scalar "knob" by which we can affect the plant. In linear control theory, we will assume that the plant is a linear filter with transfer function $\tilde{G}(\omega)$.
- The **controller** is a system that analyzes the state of the plant and implements a control procedure to the plant input. Again, we will treat this as a linear filter with transfer function $\tilde{K}(\omega)$.
- The **goal** of the control is to make the **output signal** $y(t)$ follow the **input signal** $r(t)$ as closely as possible. The control system should, however, be robust to environmental perturbations, which are something like random changes in $r(t)$. In temperature stabilization, for example, the temperature control should be robust to fluctuations in the surrounding temperature (e.g., due to the day/night cycle).

[1] For a more complete, readable introduction to control theory, see John Bechhoefer, "Feedback for physicists: A tutorial essay on control," *Reviews of Modern Physics* **77**, 783 (2005) (doi: http://dx.doi.org/10.1103/RevModPhys.77.783), available at `http://www.sfu.ca/chaos/papers/2005/rmp_reprint05.pdf`.

- The **error** is defined as the difference between the desired and actual states:

$$e(t) := r(t) - y(t). \tag{8.1}$$

  Because of the negative sign on $y(t)$ here, feeding this error signal into the controller amounts to negative feedback to the plant, assuming the low-frequency transfer characteristics of the controller and plant have no phase shift.

- The **feedback signal** $u(t)$ is the error $e(t)$, modified by the controller in frequency space.

Of course, more realistic systems may not be linear, and may have vector inputs and outputs. The scalar case is still important both conceptually and practically, so we will focus on only this here; however, note that the vector case can sometimes be treated well enough as several, parallel scalar loops—the nature of feedback control is to correct for errors, and so it can often tolerate some slop in the model. The nonlinear case is more complex in theory, but often a simple, pragmatic approach is to approximate the nonlinear system by a linearized verison, so that linear theory applies. Again, this can sometimes work even when the approximation is quite drastic.

8.2 Example: First-Order Plant, Proportional Control

As a simple example, suppose we take the plant to be a first-order, low-pass filter, with

$$\tilde{G}(\omega) = \frac{G_0}{1 - i\omega/\omega_0}, \tag{8.2}$$

where G_0 is the dc gain, and ω_0 is the cutoff frequency of the filter. A more concrete example where this model applies is temperature control of a room, where the input is a simple, electric heater, and the output is the room temperature. The low-pass-filter nature of the room is apparent in the exponential settling of the room temperature when the input (power setting of the heater, *not* a thermostat) changes.

For the controller we will implement simple **proportional control**, which just means that the control signal is proportional to the error. That is, we have a constant transfer function

$$\tilde{K}(\omega) = K_{\text{P}}, \tag{8.3}$$
(proportional controller transfer function)

where K_{P} is the **proportional gain**.

8.2.0.1 General Result: Closed-Loop Transfer Function

To analyze our simple example, we will first examine a more generally useful result. To introduce some notation, for time domain quantities like $y(t)$, we will denote their frequency-domain counterparts by $\tilde{y}(\omega)$—that is $\tilde{y}(\omega)$ is the amplitude of the frequency ω that is present in $y(t)$. Given the connections in the circuit above, we have

$$\tilde{y}(\omega) = \tilde{K}(\omega)\,\tilde{G}(\omega)\,\tilde{e}(\omega), \tag{8.4}$$

where

$$\tilde{e}(\omega) = \tilde{r}(\omega) - \tilde{y}(\omega). \tag{8.5}$$

Eliminating the error $\tilde{e}$, we have

$$\tilde{y}(\omega) = \tilde{K}(\omega)\,\tilde{G}(\omega)\big[\tilde{r}(\omega) - \tilde{y}(\omega)\big]. \tag{8.6}$$

Then

$$\tilde{y}(\omega)\big[1 + \tilde{K}(\omega)\,\tilde{G}(\omega)\big] = \tilde{K}(\omega)\,\tilde{G}(\omega)\,\tilde{r}(\omega), \tag{8.7}$$

so

$$\tilde{y}(\omega) = \frac{\tilde{K}(\omega)\,\tilde{G}(\omega)}{1 + \tilde{K}(\omega)\,\tilde{G}(\omega)}\,\tilde{r}(\omega). \tag{8.8}$$

This gives the output response $\tilde{y}$ in terms of the input $\tilde{r}$. Since the control system is linear, we have derived the transfer function for the entire control system, or the **closed-loop transfer function**

$$\tilde{T}(\omega) = \frac{\tilde{K}(\omega)\,\tilde{G}(\omega)}{1 + \tilde{K}(\omega)\,\tilde{G}(\omega)}. \qquad (8.9)$$
(closed-loop transfer function)

Relating this back to op-amps, in the limit where the gain product $\tilde{K}\tilde{G}$ becomes large, the transfer function approaches unity (and is otherwise less than unity).

8.2.1 Frequency-Domain Solution of the Example

Returning to our example, we have $\tilde{K} = K_{\text{P}}$ and $\tilde{G}$ defined by Eq. (8.2), so the closed-loop transfer function becomes

$$\tilde{T}(\omega) = \frac{K_{\text{P}}\,G_0}{K_{\text{P}}\,G_0 + 1 - i\omega/\omega_0}. = \left(\frac{K_{\text{P}}\,G_0}{K_{\text{P}}\,G_0 + 1}\right)\frac{1}{1 - i\omega/\omega_0(K_{\text{P}}\,G_0 + 1)}. \qquad (8.10)$$

This is *still* the transfer function for a low-pass filter, but now the dc gain is

$$\tilde{T}(\omega = 0) = \frac{K_{\text{P}}\,G_0}{K_{\text{P}}\,G_0 + 1}, \qquad (8.11)$$

compared to the original dc gain of G_0, and the control becomes ideal as $K_{\text{P}} \longrightarrow \infty$ (at least in this simple model; this is not true in general for *real* control systems). Also, the new cutoff frequency is $\omega_0(K_{\text{P}}G_0 + 1)$, which is larger than the original ω_0, particularly for large K_{P}. Since the cutoff frequency is inversely proportional to the decay time, we can see that a larger cutoff frequency is desirable, as it means the control system "settles" more quickly.

8.2.2 Time-Domain Solution of the Example

To examine this settling behavior more, we can also transfer the analysis for this example into the time domain. Returning to Eq. (8.6), and putting in Eq. (8.2),

$$\tilde{y}(\omega) = \tilde{K}(\omega)\,\tilde{G}(\omega)\big[\tilde{r}(\omega) - \tilde{y}(\omega)\big] = \frac{K_{\text{P}}\,G_0}{1 - i\omega/\omega_0}\big[\tilde{r}(\omega) - \tilde{y}(\omega)\big]. \qquad (8.12)$$

Rearranging a bit, we find

$$\tilde{y}(\omega)\,(\omega_0 - i\omega) = \omega_0 K_{\text{P}} G_0\big[\tilde{r}(\omega) - \tilde{y}(\omega)\big]. \qquad (8.13)$$

We can change this to the time domain by identifying the time-domain counterparts to each variable, and using $\partial/\partial t \equiv -i\omega$, to find

$$\omega_0 y(t) + \dot{y}(t) = \omega_0 K_{\text{P}} G_0\big[r(t) - y(t)\big]. \qquad (8.14)$$

Solving for $\dot{y}$,

$$\dot{y} = -\omega_0 y + \omega_0 K_{\text{P}} G_0 (r - y). \qquad (8.15)$$

There are two terms here. The first is a simple damping term, again with a time constant of $1/\omega_0$. The second term is a forcing term, where the system is "driven" by the error $e = r - y$. The system always tries to eliminate the error. Since it does so via simple exponential relaxation, it is always *stable*—it never "runs away" from the zero-error point. Note that the drive is stronger for larger K_{P}, meaning that more control has more effect on the system, as we should expect.

8.2.3 Constant Input and Proportional Droop

As a simpler version of this example, let's try a constant input $r(t) = r$. What is the steady-state solution ($\dot{y} = 0$)? From Eq. (8.15), we have

$$\omega_0 y_{\text{ss}} = \omega_0 K_{\text{P}} G_0 (r - y_{\text{ss}}), \qquad (8.16)$$

and solving for the steady-state output y_{ss}, we find

$$y_{ss} = \frac{K_P G_0}{1 + \omega_0 K_P G_0}\, r. \tag{8.17}$$

Then we see that the (proportional) control system only achieves the goal perfectly (eventually) in the limit $K_P \longrightarrow \infty$. This is the fundamental problem with proportional control: the controller only acts if there is error, so there *must* be some steady-state error, or **droop**, for any finite proportional gain K_P.[2] In op-amp circuits, the gain is large for just this reason, but for more complex, real-world control systems (electronic, mechanical, etc.), there are usually limits on K_P to maintain loop stability.

8.3 Integral Control

One approach to fixing the problem is to introduce an infinite gain *only at dc*, where the time delays that usually cause feedback-loop stability problems won't matter much. This is precisely what an integrator does: recall that an op-amp integrator (Section 7.4.2) has a gain of the form $-i/\omega RC$. More generally, integral control has a transfer function of the form

$$\tilde{K}(\omega) = \frac{iK_I}{\omega\tau}, \tag{8.18}$$
(integral controller transfer function)

where τ is a time constant and K_I is the (dimensionless) integral gain (note that τ just acts like another gain parameter here). Then noting that $1/(-i\omega)$ is an antiderivative, the controller output is

$$u(t) = \frac{K_I}{\tau}\int_0^t dt'\, e(t'). \tag{8.19}$$

That is, the controller has a built-in "memory" of past error in the feedback. This allows correction of the droop, because we no longer require an error at the *present* moment to have a nonzero control signal $u(t)$.

8.3.1 Example: First-Order Plant, Integral Control

Now back to the example that we introduced in Section 8.2. In the time domain, the same steps leading up to (8.15) now give

$$\dot{y} = -\frac{y}{\tau} + \frac{K_P G_0}{\tau^2}\int_0^t dt'\,\big[r - y(t')\big], \tag{8.20}$$

where we are taking $\tau = 1/\omega_0$ and we are still assuming a constant control input r. It is more convenient to handle an ordinary differential equation, rather than an integro-differential equation, so we can differentiate this equation to obtain

$$\ddot{y} = -\frac{\dot{y}}{\tau} + \frac{K_P G_0}{\tau^2}\big[r - y(t)\big]. \tag{8.21}$$

In steady state, $\ddot{y} = \dot{y} = 0$, and so we have

$$y_{ss} = r, \tag{8.22}$$

which means that we obtain exactly the target in steady state: there is no droop with (ideal) integral control.

8.3.2 Frequency Domain

In the frequency domain, for this example with general closed-loop transfer function (8.9) and example plant function (8.2), we have

$$\tilde{T}(\omega) = \frac{\tilde{K}(\omega)\,\tilde{G}(\omega)}{1 + \tilde{K}(\omega)\,\tilde{G}(\omega)} = \frac{1}{1 - i\omega/\omega_0 K_I - \omega^2/\omega_0^2 K_I}. \tag{8.23}$$

[2] For a good story of proportional droop, see the introduction to David Sellars, "An Overview of Proportional plus Integral plus Derivative Control and Suggestions for Its Successful Application and Implementation," `http://hdl.handle.net/1969.1/5215`.

This transfer function is second order in the denominator because of the frequency dependence of the integrator, and is qualitatively different than the first-order low-pass filter that we obtained for proportional control.

To see this, compare this to a damped, forced harmonic oscillator, which has the form

$$\ddot{y} + \gamma\dot{y} + \omega_0^2 y = f(t), \tag{8.24}$$

where γ is a damping rate, and $f(t)$ is a forcing function (we have set the mass to 1). In the frequency domain, this becomes

$$\big(-\omega^2 - i\gamma\omega + \omega_0^2\big)\tilde{y} = \tilde{f}(\omega), \tag{8.25}$$

and solving for $\tilde{y}$ gives

$$\tilde{y} = \frac{\tilde{f}(\omega)/\omega_0^2}{1 - i\gamma\omega/\omega_0^2 - \omega^2/\omega_0^2}. \tag{8.26}$$

The transfer function here has the same form as for the integrator control of a low-pass filter in Eq. (8.23). In the integrator control problem, a large integral gain K_{I} is equivalent to a large oscillation frequency ω_0 relative to the damping rate γ in the harmonic-oscillator problem. This leads to underdamped oscillations, which means the controller is overshooting the target state.

Again, note that $\tilde{T}(\omega) \longrightarrow 1$ as $\omega \longrightarrow 0$. This means that there is no steady-state droop in this integral-control example.

8.4 Proportional–Integral (PI) Control

It is, of course, possible to combine the benefits of proportional and integral control by using a controller with both features. The simplest way to combine these is a simple linear combination:

$$\tilde{K}(\omega) = K_{\text{P}} + \frac{iK_{\text{I}}}{\omega\tau}. \tag{8.27}$$
(PI-controller transfer function)

This is the transfer function for **proportional–integral (PI) control**. The second term gives integral control, which eliminates droop issues. The first term is a proportional term, which gives more high-frequency response, and thus faster setting. In the example of the single-pole plant from Section 8.2, if we work out the closed-loop transfer function, we obtain

$$\tilde{T}(\omega) = \frac{1 - i(K_{\text{P}}/K_{\text{I}})(\omega/\omega_0)}{1 - i(1+K_{\text{P}})\omega/\omega_0 K_{\text{I}} - \omega^2/\omega_0^2 K_{\text{I}}}. \tag{8.28}$$

(Again $\omega_0 = 1/\tau$ here.) Note that in the dc limit, $\tilde{T}(\omega \longrightarrow 0) = 1$, which means there is no steady-state droop, as in the integral-control case. In the high-frequency limit,

$$\tilde{T}(\omega \longrightarrow \infty) \sim \frac{-i(K_{\text{P}}/K_{\text{I}})}{(1+K_{\text{P}})/K_{\text{I}} - i\omega/\omega_0 K_{\text{I}}} = \frac{-iK_{\text{P}}}{(1+K_{\text{P}}) - i\omega/\omega_0}. \tag{8.29}$$

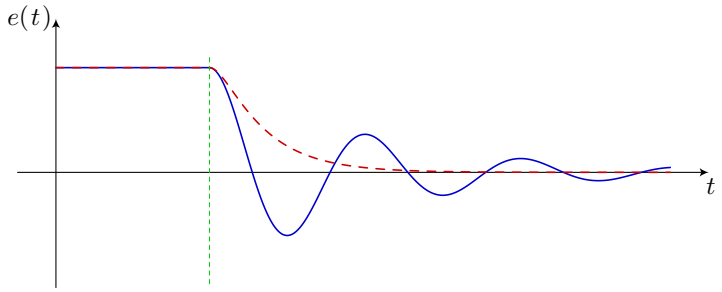
In the high-frequency limit, the transfer function reduces to a first-order transfer function, in which case we no longer expect overshoot behavior, as we did in the integral case.

8.5 Proportional–Integral–Derivative (PID) Control

In **proportional–integral–derivative (PID) control**, the idea is to add a derivative term to PI control, so the transfer function is

$$\tilde{K}(\omega) = K_{\text{P}} + \frac{iK_{\text{I}}}{\omega\tau} - i\omega\tau K_{\text{D}}. \tag{8.30}$$
(PID-controller transfer function)

Here, K_{D} is the derivative gain. Intuitively, this can help in cases where overshoot and ringing is a problem. Qualitatively, consider the overshooting case below in blue, where the input is suddenly changed at the time marked by the dashed green line.



Qualitatively, the overshoot occurs because the slope of the error signal is too steep, due to the action of the controller. By putting in a term proportional to the derivative, the controller "senses" the steep slope corresponding to an impending overshoot, and reduces the control action. This can result in better settling, as in the dashed red line. In terms of the closed-loop transfer function, the effect of the derivative gain is to modify the damping coefficient of the feedback system, which can eliminate the ringing and promote better settling.

Setting the parameters for a PI or PID loop is something of an art. We won't get into this here, but one reasonably simple method for setting the gain parameters is the **Ziegler-Nichols method**.[3]

[3] J. G. Ziegler and N. B. Nichols, "Optimum Settings for Automatic Controllers," *Transactions of the ASME* **64**, 759 (1942), copies available at `http://chem.engr.utc.edu/Student-files/x2008-Fa/435-Blue/1942-paper.pdf` and `http://www.driedger.ca/Z-N/Z-n.pdf`. See also Allard Mosk, "Tutorial on Experimental Physics of Ultracold Gases," in *Interactions in Ultracold Gases: From Atoms to Molecules*, Matthias Weidemüller and Claus Zimmerman, Eds. (Wiley-VCH, 2003), p. 215 (doi: 10.1002/3527603417.ch5).

Part II

Digital Electronics

Chapter 9

Binary Logic and Logic Gates

9.1 Binary Logic

The idea behind *binary logic* is to represent information using only two states. You can call these states `TRUE` and `FALSE`, or you can use the corresponding numerical values 1 and 0. We will explore in much more detail how to represent and use information in this form, but for now, note that we call the fundamental (abstract) element that carries these states a **bit**. That is, a single bit can have either the values 0 or the value 1.

The idea behind **digital logic** and **digital electronics** is to represent the binary states by two different electronic states, usually different voltages or voltage ranges, but sometimes different currents. For example, the standard for **transistor-transistor logic (TTL)** is to use nominal voltages of 0 V for `FALSE`, and +5 V for `TRUE`. We will get into more detailed specifics later.

Changing information into a digital representation has advantages and disadvantages. The main *disadvantage* of this approach is that it is necessary to **sample** analog signals (i.e., change continuous signals into discrete representations). The main *advantage* is in robustness to noise, as long as the noise amplitude is far below the physical separation between the logic states (e.g., TTL logic is robust to noise interference provided the noise is smaller than 5 V). Of course, for sophisticated logic systems (computers), often the advantages far outweigh the disadvantages.

We will treat binary logic as an abstract concept for now, and learn how to manipulate binary information. Then we will come back later to the physical implementation of binary logic.

9.2 Binary Arithmetic

In *binary arithmetic*—the binary analogue of the more usual arithmetic—the first thing to deal with is how to *represent* numbers in binary. To keep things relatively simple, we will stick to representing *integers* in binary (as opposed to rational approximations to real numbers, which are represented in either **fixed-point** or **floating-point** notation, the latter of which is more complicated).

9.2.1 Unsigned Integers

The most basic form of a binary integer is an **unsigned integer**. Unsigned integers are just like decimal integers, but instead of counting from 0–9 and then carrying a 1 to the next place, you just count from 0-1 and then carry instead. (So the counting is 0, 1, 10, 11, 100, 101, 110, 111, ...) You can understand converting between binary and decimal best via an example. Suppose that we have the unsigned integer 1011_2 (the subscript "2" denotes binary, or base-2 arithmetic). There are four digits, which represent, from right to left, the "ones," "twos," "fours," and "eights" places (just like the ones, tens, etc. in decimal counting). Then proceeding from the ones (rightmost) place, or the **least-significant bit (LSB)**, to the eights (leftmost)

place, or the **most-significant bit (MSB)**.

$$1011_2 = 1 \times 2^0 + 1 \times 2^1 + 0 \times 2^2 + 1 \times 2^3 = 1 + 2 + 8 = 11. \tag{9.1}$$

Note that with N digits, we can represent 2 values with each bit, for a total of 2^N numbers (i.e., ranging from 0 to $2^N - 1$).

9.2.1.1 Binary-Coded Decimal

Another representation of unsigned integers comes in **binary-coded decimal (BCD)**, where the idea is to convert each digit in a decimal number to binary, using 4 bits per decimal number. For example, $11_{10} = 1011_2$, but in BCD, this would be written 00010001. This representation is "wasteful" in that there are 16 states for each digit but only 10 decimal digits, but this representation is very convenient for implementations of digital numeric displays.

9.2.1.2 Hexadecimal

Hexadecimal arithmetic is just base-16 arithmetic. The 16 states are represented by 0–9 as usual, and the "extras" by A–F for the values 10–15. Since 4 bits, or a **nybble** (8 bits is a **byte**) has 16 states, a single hexadecimal digit is a convenient and compact representation for a binary nybble. Thus, for example, $10111010_2 = BA_{16}$.

9.2.2 Negative Values and Sign Conventions

Besides unsigned integers, it is useful to represent *negative* integers in binary. There are multiple conventions for this, however.

9.2.2.1 Sign-Magnitude Convention

The simplest convention, the **sign-magnitude convention**, is to tack on an extra bit (as a new MSB) to represent the sign, and the rest of the digits are just like an unsigned integer. For example, one nybble ranges from 0–15 as an unsigned integer, but as a signed value, it ranges from -7 to $+7$ as a signed integer (the three LSB's range from 0–7, and the MSB gives the sign). Note that one value (1000) is "wasted" in this convention, because it is not different from (0000). The main advantage is the simplicity of the scheme. You can see the relatively serious disadvantage, however, by considering a couple of example numbers,

$$0001_2 = 1_{10}, \qquad 1001_2 = -1_{10}. \tag{9.2}$$

Unfortunately, adding these two numbers gives $1010_2 = -2_{10}$, but really we'd like these to add to *zero*.

9.2.2.2 Two's Complement

Preserving this additive-inverse property of negative numbers is the idea behind the **2's-complement** representation: if n is a positive integer, just define the number $(-n)$ such that it satisfies $n + (-n) = 0$ in binary addition. For example, suppose we have

$$n = 0001_2 = 1_{10}. \tag{9.3}$$

Then -1_{10} in 2's-complement notation is

$$-n = 1111_2 = -1_{10}. \tag{9.4}$$

To see this, first note that

$$n + (-n) = 10000_2, \tag{9.5}$$

but the important point is that we *drop* the MSB, because we regard addition in 4-bit binary as being modulo 16.

The advantage, of course, is that addition works as expected with positive and negative numbers in 2's-complement notation, and so does multiplication.

There are a couple of useful procedures for finding 2's-complement values (i.e., for finding the negative counterpart of a positive number):

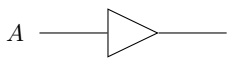
1. Start by exchanging $0 \longleftrightarrow 1$ on each digit, then add 1 to the result. For example, in the example above,
$$\texttt{0001}_2 \longrightarrow \texttt{1000}_2 + 1 = \texttt{1111}_2. \tag{9.6}$$
2. Note also that in N-bit arithmetic, $2^{N-1} = -2^{N-1}$. So we just need to figure out what number to add to -2^{N-1} to get the number we want. For example, in the above example in 4-bit arithmetic, $2^3 = \texttt{1000}_2 = -2^3 = -8$. We want -1, which means we have to add 7 to -8. Since $7 = \texttt{0111}_2$, we just say
$$-1 = -8 + 7 = \texttt{1000}_2 + \texttt{0111}_2 = \texttt{1111}_2. \tag{9.7}$$
Alternately, we are just saying that for any negative number that we want in N-bit arithmetic, add 2^N and then find the *unsigned* binary value. (In the example, add 16 to -1 to get 15, or $\texttt{1111}_2$. In this convention, we're just taking the unsigned range of $2^{N-1}+1$ to 2^N, and shifting the whole block to below zero.

9.3 Logic Gates

So far, we have discussed only binary-logic values and how to use them to represent numbers. But we also need to implement *transformations* on logic values, which are accomplished via **logic gates**, which is basically a logic-valued function of logic variables. We will talk about the simplest logic gates now, and just mention that more complicated gates can be represented in terms of the simpler ones.

9.3.1 One-Input Gates

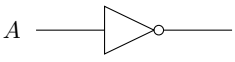
The simplest logic gates are the one-input gate, which takes only one logic value as input. A simple example is the **buffer gate**, which simply copies its input A to the output Q. The symbol for the buffer is below.



A	$Q = A$
1	1
0	0

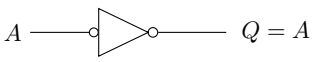
Above on the right is the **truth table**, a table enumerating all inputs and the corresponding output values.

The other main one-input gate is the **inverter** or `NOT` gate, which changes the state of the input. The symbol and truth table are below.

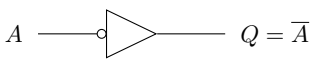


A	$Q = \overline{A}$
1	0
0	1

Note the circle "$\circ$" in the diagram represents a `NOT` operation, which is denoted symbolically by a bar (that is, if $A = 1$, the $\overline{A} = 0$). This same `NOT` operation may be applied to inputs as well. For example, this is a buffer gate,



and this is another `NOT` gate.

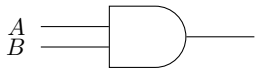


The buffer and `NOT` gates are the only one-input gates. The only other possibilites (in terms of truth-table content) have a fixed output for any input, which is usually not drawn as a gate with an input.

9.3.2 Two-Input Gates

9.3.2.1 AND and NAND

Two-input gates are important, and easily available as electronic components. The first gate is the `AND` gate, whose symbol and truth table are shown below.

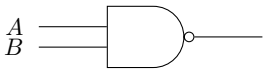


$Q = A \cdot B = AB = A \wedge B$

A	B	$Q = AB$
1	1	1
1	0	0
0	1	0
0	0	0

There are several notations for the `AND` operation in the diagram. Note that the output is only `TRUE` if *both* inputs are `TRUE`.

Adding a `NOT` to the output of the `AND` gate gives a `NAND` gate (i.e., `NOT AND`), which is just the negation of the `AND` gate.



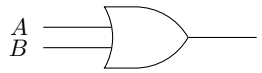
$Q = \overline{A \cdot B} = \overline{AB} = \overline{A \wedge B}$

A	B	$Q = \overline{AB}$
1	1	0
1	0	1
0	1	1
0	0	1

This gate is more important than it may seem at first glance, as we'll return to below.

9.3.2.2 OR and NOR

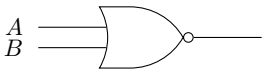
The next gate is the `OR` gate, whose operation is symbolically represented by "+."



$Q = A + B = A \vee B$

A	B	$Q = A + B$
1	1	1
1	0	1
0	1	1
0	0	0

The output here is `TRUE` if either input is `TRUE` (or both inputs are `TRUE`). Of course, we can add a `NOT` to the output.



$Q = \overline{A + B} = \overline{A \vee B}$

A	B	$Q = \overline{A + B}$
1	1	0
1	0	0
0	1	0
0	0	1

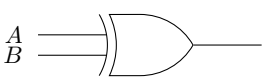
Thus, we obtain the `NOR` gate (`NOT OR`).

9.3.2.3 Universal Gates

The `NAND` and `NOR` gates are special, because they are **universal gates**. That is, *any* logic operation can be realized by connecting a bunch of `NAND` gates, or by connecting a bunch of `NOR` gates.

9.3.2.4 XOR and XNOR

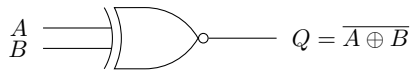
We'll briefly also mention the `XOR` gate ("exclusive `OR`"),



$Q = A \oplus B$

A	B	$Q = A \oplus B$
1	1	0
1	0	1
0	1	1
0	0	0

which is just like the OR, but the output is FALSE if *both* inputs are TRUE. The complement of the XOR gate is the XNOR gate ("exclusive NOR"), which is again like the NOR gate except for the case of two TRUE inputs.

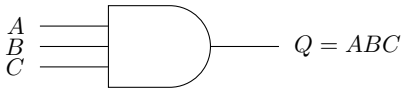


A	B	$Q = \overline{A \oplus B}$
1	1	1
1	0	0
0	1	0
0	0	1

In mathematical logic, the XNOR is the same as "if and only if."

9.3.3 More Complex Gates

More complex gates are possible; for example, consider the 3-input AND gate below.



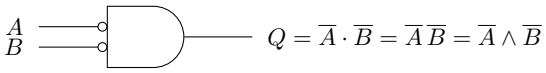
A	B	C	$Q = ABC$
1	1	1	1
1	1	0	0
1	0	1	0
1	0	0	0
0	1	1	0
0	1	0	0
0	0	1	0
0	0	0	0

The idea is a reasonable obvious generalization of the 2-input AND: the output is TRUE only when *all* inputs are TRUE.

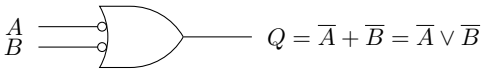
9.4 Circuit Practice

Here are a couple of gates with negated inputs.

(a) Work out the truth table and find which 2-input gate that we introduced above is equivalent.



(b) Do the same for this gate.



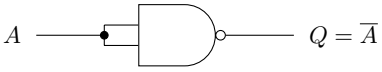
(c) How do you make an inverter from a NAND gate?

Solution.

(a) NOR gate.

(b) NAND gate.

(c) Tie the inputs together.



A	$Q = \overline{A}$
1	0
0	1

9.5 Exercises

Problem 9.1

Convert:

(a) 89_{10} to 8-bit, unsigned binary

(b) -89_{10} to 8-bit, signed (sign-magnitude) binary

(c) -89_{10} to 8-bit, signed (2's complement) binary

(d) 89_{10} to hexadecimal

(e) $\texttt{ABCD}_{16}$ to decimal

(f) $\texttt{0110011001100110}_2$ to hex

(g) $\texttt{0110011001100110}_2$ to decimal

(h) $\texttt{10011001}_2$ (2's complement binary) to decimal

(i) $\texttt{10011001}_2$ (sign-magnitude binary) to decimal

(j) $\texttt{11111111}_2$ (unsigned) to decimal

Problem 9.2

Convert:

(a) 75_{10} to 8-bit, unsigned binary

(b) -75_{10} to 8-bit, signed (sign-magnitude) binary

(c) -75_{10} to 8-bit, signed (2's complement) binary

(d) 75_{10} to hexadecimal

(e) $\texttt{ABBA}_{16}$ to decimal

(f) $\texttt{1010101010101010}_2$ to hex

(g) $\texttt{1010101010101010}_2$ to decimal

(h) $\texttt{11011101}_2$ (2's complement binary) to decimal

(i) $\texttt{11011101}_2$ (sign-magnitude binary) to decimal

(j) $\texttt{11111101}_2$ (unsigned) to decimal

Problem 9.3

(a) Suppose x is a power of 2 (i.e., $x = 2^n$ for some positive integer n, $n \in \mathbb{Z}^+$). What does x "look" like when written out in (unsigned) binary? (That is, how can you recognize powers of two, when written in binary, just by looking at them?)

(b) In writing computer programs it is sometimes useful to check whether an integer is a power of 2. (One example is in computing numerical Fourier transforms, where the most common algorithms operate only on arrays whose lengths are powers of 2.)

A nice trick for checking if x is a power of 2 is to compute $x \wedge (x-1)$. That is, subtract 1, and compute the bitwise `AND` with the original. How do you tell from the result if x is a power of 2? (You might try this on some examples to see the pattern.)

Note: "bitwise `AND`" means to compute the `AND` of corresponding binary digits. For example, $\texttt{1100}_2 \wedge \texttt{1010}_2 \equiv \texttt{1100}_2 \cdot \texttt{1010}_2 = \texttt{1000}_2$.

Problem 9.4

Short-question potpourri:
(a) For any integer expressed in decimal, suppose you add the digits. The result is divisible by 3 if and only if the original number is also divisible by 3. Does this statement also hold for binary numbers? If so, explain why. If not, provide a counterexample.

(b) What is $\texttt{1000 0000}$ in decimal? Interpret the given number as an 8-bit, signed (2's complement) binary number.

(c) What is $\texttt{111}_{10}$ in unsigned binary?

(d) What is $\texttt{111}_{10}$ in hex?

(e) Suppose $\texttt{1010}_2 \oplus B = \texttt{1100}_2$, where the operation is bitwise. What is B in decimal?

Problem 9.5

Short-question potpourri:
(a) What is $2^{38\,457} - 3$ in (unsigned) binary? (I suggest describing *how* to write down the binary expression, not actually writing it out.)

(b) What is $\texttt{1010 1010}_2 - \texttt{0101 0101}_2$? (Do the calculation in binary.)

(c) What is $\texttt{1101}_2$ in hex?

(d) What is $\texttt{1101}_2 + \texttt{3}_{16}$? Interpret both numbers as 4-bit, 2's complement binary numbers, and give your answer in 4-bit, 2's complement binary.

(e) Suppose you generalize decimal-fraction notation (e.g., $0.9 = {}^{9}/_{10}$) to binary fractions in the obvious way (e.g., $\texttt{0.1}_2 = {}^{1}/_{2}$). What is the value of $\texttt{0.}\overline{\texttt{1}}_2 \equiv \texttt{0.111}\ldots_2$ (i.e., the overbar here means a repeating digit) in decimal?

Chapter 10

Boolean Algebra

10.1 Algebras and Boolean Algebra

Intuitively, a **Boolean algebra** is an abstract, compact notation for logic (in which we will see that $1+1 = 1$). A Boolean algebra is defined on the set $\{0, 1\}$ (these are the "values" for Boolean variables), and has two **binary operations** "$+$" and "$\cdot$" defined, though not the usual addition and multiplication. The $+$ operation is defined by the truth table for the `OR` gate, while the $\cdot$ operation is defined by the truth table for the `AND` gate. Recall that the truth tables for the `AND` and `OR` operations on Boolean variables A and B are invariant under the exchange of A and B, so both $+$ and $\cdot$ are **commutative** (i.e., the order of the variables don't matter):

$$A + B = B + A, \qquad AB = BA. \tag{10.1}$$

(We're not bothering to write the "$\cdot$" explicitly here.) These operations are also **associative**, which means that the order of two successive operations does not matter:

$$A + (B + C) = (A + B) + C, \qquad A(BC) = (AB)C. \tag{10.2}$$

The other usual algebraic property that holds here is the **distributive** property:

$$A(B + C) = (AB) + (AC). \tag{10.3}$$

Finally, a number of simple identities hold for the Boolean binary operators:

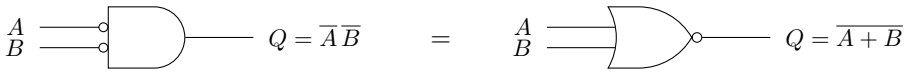
1. $\overline{\overline{A}} = A$
2. $A \cdot 0 = 0$
3. $A \cdot 1 = A$
4. $A \cdot A = A$
5. $A \cdot \overline{A} = 0$
6. $A + 0 = A$
7. $A + 1 = 1$
8. $A + A = A$
9. $A + \overline{A} = 1$

Roughly speaking, the `NOT` operation (bar) is something like a minus sign, in which case some of these identities seem familiar, but some seem less so (like $A + 1 = 1$).

10.2 Boolean-Algebraic Theorems and Manipulations

10.2.1 De Morgan's Theorems

Recall the circuit practice from Section 9.4, where we examined `AND` and `OR` gates, where both inputs are negated. For example, the negated-input `AND` gate is equivalent to `NOR` gate, as shown schematically below.



In algebraic notation, this is

$$\overline{A} \cdot \overline{B} = \overline{A + B}. \tag{10.4}$$
(De Morgan theorem)

Similarly

$$\overline{A} + \overline{B} = \overline{A \cdot B}. \tag{10.5}$$
(De Morgan theorem)

These are extremely useful in transforming negated expressions, as we will see.

10.2.2 Absorption Theorems

Two other useful theorems are called **absorption theorems**:

$$\begin{aligned} A + (A \cdot B) &= A \\ A \cdot (A + B) &= A. \end{aligned} \tag{10.6}$$
(absorption theorems)

We will leave the proofs as exercises (in circuit practice).

10.2.3 Another Theorem

Here is another theorem that is often useful:

$$A + \overline{A}B = A + B. \tag{10.7}$$
(negated AND theorem)

We will again leave the proof as an exercise, but essentially this is saying that because of the `OR` with A, the $\overline{A}$ never really matters.

10.2.4 Example: XOR Gate

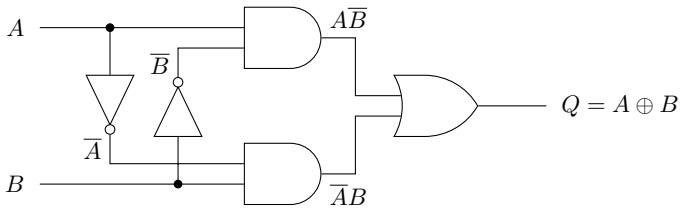
As an example of Boolean algebra and implementation of algebraic expressions in gates, consider the `XOR` operation, where we would like to show that

$$A \oplus B = \overline{A}B + A\overline{B}. \tag{10.8}$$
(`XOR` expression)

We can first do this by working through the truth table for the right-hand side, and verifying that it matches the truth-table results for $A \oplus B$.

A	B	$\overline{A}B$	$A\overline{B}$	$A \oplus B$
1	1	0	0	0
1	0	0	1	1
0	1	1	0	1
0	0	0	0	0

Now using this expression, we can show how to implement an *XOR* gate, in terms of regular gates.



To trace through this, the negations $\overline{A}$ and $\overline{B}$ are realized with `NOT` gates, and finally two `AND` gates and an `OR` gate to generate the correct combination.

10.2.4.1 NAND-Gate Realization

As we alluded to before, `NAND` and `NOT` gates are universal, and can be used to realize any gate. So how can we realize an `XOR` gate out of only `NAND` gates, for example? Let's do some algebraic transformations to see how to do this. First, starting with the expression (10.8),

$$A \oplus B = \overline{A}B + A\overline{B}, \tag{10.9}$$

we can add in $A\overline{A} = 0$ and $B\overline{B} = 0$ to obtain

$$A \oplus B = B(\overline{A} + \overline{B}) + A(\overline{A} + \overline{B}). \tag{10.10}$$

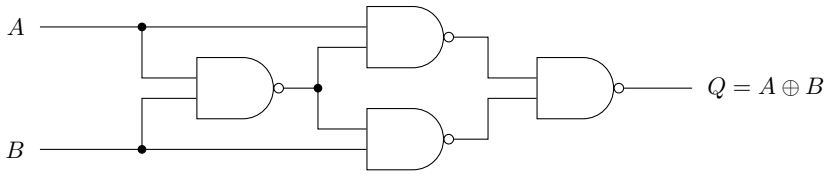
Then using the second De Morgan theorem (10.5), $\overline{A} + \overline{B} = \overline{A \cdot B}$,

$$A \oplus B = B(\overline{AB}) + A(\overline{AB}). \tag{10.11}$$

Using the same theorem once more,

$$A \oplus B = \overline{\overline{[B(\overline{AB})]}\,\overline{[A(\overline{AB})]}}. \tag{10.12}$$

Now notice that every operation here is a `NAND`, and we need one operation to generate $\overline{AB}$, two more to combine it with A and B, and one more for the final combination. The circuit realizing this expression is shown below.



Note that we can obtain a simpler but less-efficient expression by applying the De Morgan theorem only once as follows:

$$A \oplus B = \overline{A}B + A\overline{B} = \overline{\overline{(\overline{A}B)}\,\overline{(A\overline{B})}}. \tag{10.13}$$

This is less efficient in terms of `NAND` gates: two `NAND`s are needed to make $\overline{A}$ and $\overline{B}$, two more to make the combinations $\overline{A}B$ and $A\overline{B}$, and one more to make the final combination (a total of 5).

10.2.5 Example: Algebraic Simplification

As another example of simplifying an expression, consider the three-variable expression $(A + B) \cdot (A + C)$. Starting out, we can distribute twice,

$$\begin{aligned}(A + B) \cdot (A + C) &= A \cdot (A + C) + B \cdot (A + C) \\ &= AA + AC + AB + BC.\end{aligned} \tag{10.14}$$

Then $AA = A$, and using the first absorption theorem in Eqs. (10.6) to write $A + AC = A$,

$$\begin{aligned}(A + B) \cdot (A + C) &= (A + AC) + AB + BC \\ &= A + AB + BC.\end{aligned} \tag{10.15}$$

Then also $A + AB = A$, so

$$(A + B) \cdot (A + C) = A + BC, \tag{10.16}$$

a somewhat simpler expression (two operations vs. three).

10.3 Karnaugh Maps

It can often be difficult to see how to realize a particular logic function in terms of logic gates, just via algebraic manipulations. One tool that makes this more intuitive, at least for small numbers of inputs (3 or 4), is the **Karnaugh map**. (The cases of 1 and 2 inputs we've already mostly covered with standard gates, and it's easy to do these exhaustively.)

The first idea behind a Karnaugh map is to make a 2D table of logic inputs and truth-table values. The second twist is to order the inputs using a 2-bit **Gray code**, which means that we count as `00`, `01`, `11`, `10`, instead of the usual binary-counting order. The point is when counting this way we change only 1 bit at a time (in regular binary, this doesn't happen when we count from `01` to `10`). The motivation for Gray codes comes from mechanical implementations of logic, where you may get spurious transitional states if bits don't change synchronously (this happens in some fast logic circuits as well). That is, when counting from `01` to `10`, the actual sequence may be `01` to `00` to `10` if the LSB changes before the MSB. In terms of the Karnaugh map, the idea is to keep "related" input states grouped together.

The process of hunting for simplifications in a Karnaugh map is hard to explain, but it's easy to get the idea by studying a few examples.

10.3.1 Three-Input Example

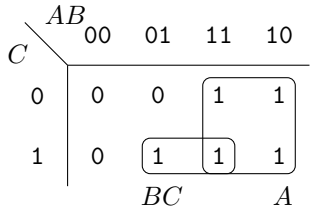
Before, as a Boolean-algebraic example, we showed in Eq. (10.16) that

$$(A + B) \cdot (A + C) = A + BC. \tag{10.17}$$

We will show how to obtain this and other transformations via the Karnaugh map. The first task is to write out the diagram as a table. Notice that the four values of AB are along the top, in Gray-coded order, and the two C values are along the side. We are also writing out each output value for each set of possible input values, so this is just a truth table in 2D form, here for $(A + B) \cdot (A + C)$.

C \ AB	00	01	11	10
0	0	0	1	1
1	0	1	1	1

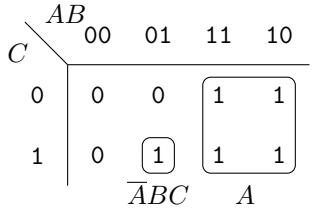
Now to analyze this, the idea is to look for blocks of `1`'s in square or rectangular shapes (2×1, 2×2, etc.). One example is below.



$A + BC$

The 2×2 block here corresponds to every input where $A = 1$, hence we have labeled it "A." Similarly, the 2×1 block correponds to inputs where $B = 1$ and $C = 1$ (hence $BC = 1$), so we label it "BC." The entire group of `1`'s is the union of these two, so the logical expression is the `OR` of these two blocks, hence an equivalent expression is $A + BC$.

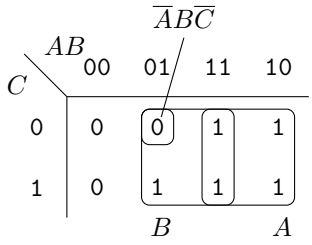
Generally speaking, bigger blocks correspond to simpler expressions, so the best simplifications occur by covering the `1`'s with large blocks. Also, usually it is best to look for blocks with dimensions of 2, 4, 8, etc. As an example, note that we could have done the last covering without any overlap if we kept the 4×4 block and then introduced a 1×1 block, as below.



$A + \overline{A}BC$

The small block corresponds to $\overline{A}BC$, since we have to restrict all three variables, and this leads to the more complicated (but equivalent) expression $A + \overline{A}BC$.

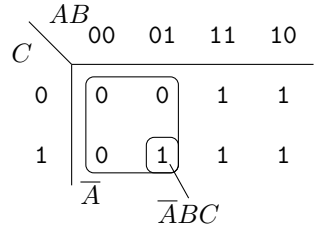
We can also look at some other attempts to simplify with a Karnaugh map that will yield less compact results, just to illustrate the technique. For example, we can "overcover" the `1`'s by using two 4×4 blocks, A and B, and then combine them. However, we must exclude one location that has a zero; the location is $\overline{A}B\overline{C}$.



$A + B\overline{\overline{A}B\overline{C}}$

Thus, to combine these, we *negate* the null block and `AND` the result with the B block to obtain $B\overline{\overline{A}B\overline{C}}$. We then `OR` this with A to obtain $A + B\overline{\overline{A}B\overline{C}}$.

Another possibility is to use a similar technique, but focusing on the `0`'s. For example, we can identify a block of mostly zeros, $\overline{A}$. However, we must exclude the 1, which is located at $\overline{A}BC$.

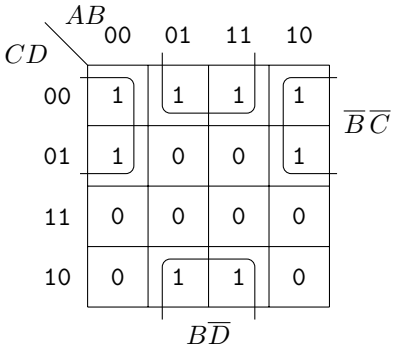


$\overline{\overline{A}\,\overline{\overline{A}BC}}$

So to find the region of 0's, we have to negate the 1 block and AND this with the 0 block, to obtain $\overline{A}\,\overline{\overline{A}BC}$. Then to get the block of 1's we must negate the overall result, to obtain $\overline{\overline{A}\,\overline{\overline{A}BC}}$. Note that by De Morgan's theorems this is equivalent to $A + \overline{A}BC$, as we showed two diagrams ago.

10.3.2 Four-Input Example

In addition to three-input problems, it is not much harder to extend the analysis to four-input problems. For example, consider the following truth table.

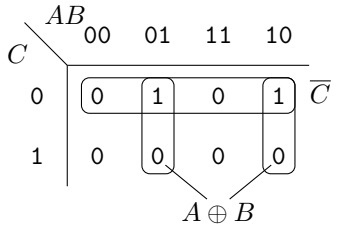


$\overline{B}\,\overline{C} + B\overline{D}$

One thing to notice is that we have extended the vertical direction to cover the two variables C and D together, and the other is that we have put in 4×4 blocks that "wrap" around from top to bottom or right to left. That is, the Karnaugh map has periodic boundary conditions for the purposes of finding blocks.

10.3.3 XOR Example

In searching for blocks, it is somewhat harder to see XOR and XNOR blocks, but it is possible. An example is below.



$(A \oplus B)\overline{C}$

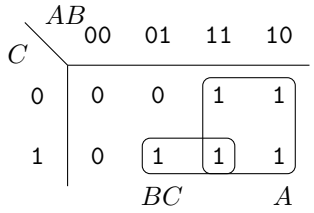
Due to the ordering of the horizontal axis, the $A \oplus B$ block is split, but we can combine it with the $\overline{C}$ via an AND (to intersect the blocks) to obtain a relatively simple expression.

Note that some flexibility is usually beneficial when using a Karnaugh map: it is not necessarily a good tool for finding solutions in terms of a *particular* gate (e.g., all NAND gates).

10.3.4 Race Hazards

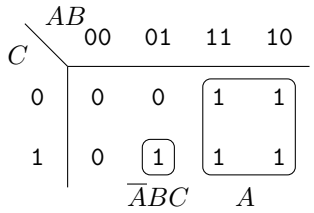
A **race condition** is a spurious output of a circuit if the inputs don't change state simultaneously (i.e., a "glitch"). This can be a big problem if this output is the input to a latch or a memory circuit that will "trigger" on the glitch.

Intuitively, in a Karnaugh map, a glitch is possible if the changing inputs cross between disjoint blocks of 1's, because the output state is being controlled by transitions of two gates feeding into the same final gate. For example, returning to the $(A + B)(A + C)$ example, suppose we make a transition between 111 to 011 in ABC. In this logic realization,



$A + BC$

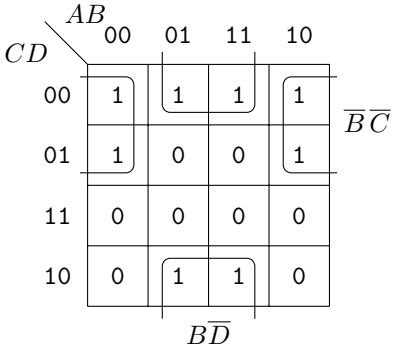
we stay inside the BC block, so we don't expect any glitches: the output stays at `1` during the transition. However, in *this* realization,



$A + \overline{A}BC$

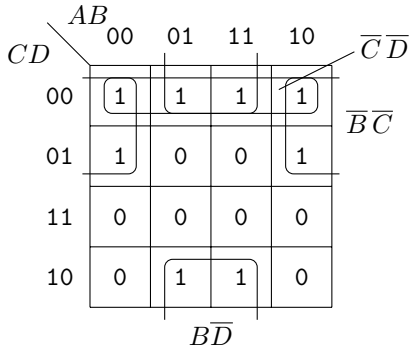
we must cross between blocks, so a glitch is possible. Specifically, when A goes from `1` $\longrightarrow$ `0`, a slight delay in $\overline{A}$ going from `0` $\longrightarrow$ `1` results in the output going momentarily to `0` during the input transition, even though it should remain as `1`.

As another example, let's return to the four-input example.



$\overline{B}\,\overline{C} + B\overline{D}$

There is a similar problem here when $ABCD$ goes from `1100` $\longrightarrow$ `1000`, because we cross in between blocks. However, by adding another block, we can "protect" the circuit from glitches in this transition. Here, we add $\overline{C}\,\overline{D}$, and combine it with an `OR` operation.



$\overline{B}\,\overline{C} + B\overline{D} + \overline{C}\,\overline{D}$

Of course, the price for robustness is a more complicated expression.

10.4 Circuit Practice

10.4.1 Boolean-Algebra Theorems

Here, you should prove two things that we only introduced earlier.

(a) Prove the first absorption theorem in Eqs. (10.6):

$$A + (A \cdot B) = A. \tag{10.18}$$

Use a truth table or algebra.
(b) Prove the second absorption theorem in Eqs. (10.6):

$$A \cdot (A + B) = A. \tag{10.19}$$

Solution.
(a) First suppose $B = 0$. Then

$$A \cdot B = A \cdot 0 = 0, \tag{10.20}$$

and so

$$A + (A \cdot B) = A + 0 = A. \tag{10.21}$$

Now take the other case, where $B = 1$. Then

$$A \cdot B = A \cdot 1 = A, \tag{10.22}$$

and so

$$A + (A \cdot B) = A + A = A. \tag{10.23}$$

(b) Using the same method, first suppose $B = 0$. Then

$$A + B = A + 0 = A, \tag{10.24}$$

and so

$$A \cdot (A + B) = A \cdot A = A. \tag{10.25}$$

Now taking the other case $B = 1$,

$$A + B = A + 1 = 1, \tag{10.26}$$

and so

$$A \cdot (A + B) = A \cdot 1 = A. \tag{10.27}$$

10.4.2 Karnaugh Map

Write down the Karnaugh map and a logic circuit for the following function: the output is `1` if and only if the input, a 3-bit unsigned integer, is prime. (Don't count 0, 1, or 2 as prime integers.)

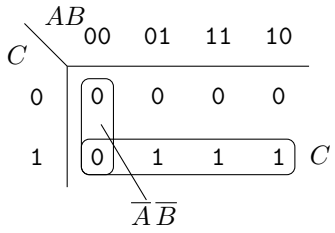
Solution. The primes are 3, 5, and 7. In binary, these are `011`, `101`, and `111`. Hence the Karnaugh map:

C \ AB	00	01	11	10
0	0	0	0	0
1	0	1	1	1

$(A + B)C$

$(A + B)C$

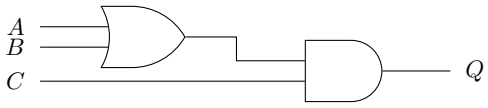
The simplest solution for a logic gate is to make a 3×1 block, noting that the AB part is specified by $(A + B)$.



$$\overline{\overline{A}\,\overline{B}C} = (A + B)C$$

An alternate, and equivalent solution (via De Morgan's theorems) is shown below, by starting with the 4×1 block C, and then excluding the 0.

The circuit to realize this function is shown below.



10.5 Exercises

Problem 10.1

Simplify the expression

$$Q = \overline{A}\,\overline{B}\,\overline{C} + \overline{A}\,\overline{B}C + \overline{A}B\overline{C} + A\overline{B}\,\overline{C}, \tag{10.28}$$

and draw a logic circuit that realizes it. (This can be done with only 3 2-input gates and 1 3-input gate; try to at least reduce this somewhat, and it's best if your solution reflects the symmetry of the original expression. Also, try to use algebraic transformations rather than writing out truth tables.)

Problem 10.2

Simplify the expression

$$Q = \overline{ABC} + \overline{A}\,\overline{B}C + \overline{A}B\overline{C} + A\overline{B}\,\overline{C}, \tag{10.29}$$

and draw a logic circuit that realizes it. (This is possible with only 1 3-input gate; try to at least reduce this somewhat, and it's best if your solution reflects the symmetry of the original expression. Also, try to use algebraic transformations rather than writing out truth tables.)

Problem 10.3

(a) Simplify the following Boolean expression:

$$Q = (A + B)(\overline{B} + A) + \overline{\overline{AB} + \overline{A + \overline{B}}} + B. \tag{10.30}$$

(b) Sketch a realization of this expression (after simplifying!) using only 2-input `NAND` gates.

Problem 10.4

(a) Simplify the following Boolean expression:

$$Q = \overline{(\overline{A} + \overline{B})(\overline{A + B}) + (A + \overline{B})(\overline{A} + B)}. \tag{10.31}$$

(b) Sketch a realization of this expression (after simplifying!) in terms of only `XNOR` gates.

Problem 10.5

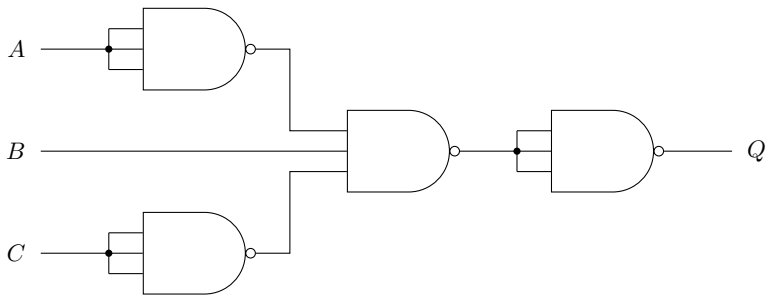
(a) Simplify the following Boolean expression:

$$Q = \overline{\left[(\overline{A}\,\overline{BC})(\overline{B}\,\overline{AC})\right]\left[(\overline{AB})(\overline{AB\overline{C}})\right]}. \tag{10.32}$$

(b) Sketch a realization of this expression (after simplifying!) in terms of only `NOR` gates.

Problem 10.6

Consider the following circuit, based on 3-input `NAND` gates.

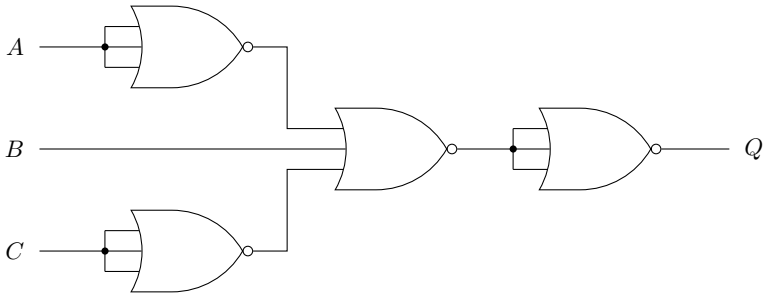


(a) Write down the logic (Boolean) expression for the circuit.

(b) Write down the truth table.

Problem 10.7

Consider the following circuit, based on 3-input `NOR` gates.



Problem 10.8

Show how you can realize an `XOR` gate ($A \oplus B = \overline{A} \cdot B + A \cdot \overline{B}$) using only `NOR` gates.

Problem 10.9

Show how you can realize an `XNOR` gate ($\overline{A \oplus B} = \overline{\overline{A} \cdot B + A \cdot \overline{B}}$) using only `NAND` gates.

Problem 10.10

Write out the Karnaugh map for a circuit where the output is true if the input (a 3-bit, unsigned integer, 0–7) is in the Fibonacci sequence. Give a circuit implementation in terms of 2-input gates.

Problem 10.11

Write out the Karnaugh map for a circuit where the output is true if the input (a 3-bit, unsigned integer, 0–7) is one of the first 6 digits of π. Give a circuit implementation in terms of 2-input gates.

Problem 10.12

A **semiprime number** is a positive integer that is the product of two prime numbers. The prime numbers need not be distinct, 1 doesn't count as one of the primes. For example, 0–3 are not semiprime, but 4 is.

(a) Write down the Karnaugh map for the function of the boolean variables A, B, C, and D, which is true when the concatenation $ABCD$ (when converted to decimal as an unsigned integer) is semiprime.

(b) Find a (reasonably simple) boolean expression for this 4-bit semiprime function you diagrammed in (a).

(c) Sketch a logic implementation of this function in terms of logic gates.

Problem 10.13

(a) Write down the Karnaugh map for the function of the Boolean variables A, B, C, and D, which is true when the concatenation $ABCD$ (when converted to decimal as an unsigned integer) is greater than or equal to 6.

(b) Find a (reasonably simple) boolean expression for the logic function you diagrammed in (a).

(c) Sketch a logic implementation of this function in terms of only 2-input `NAND` gates.

Problem 10.14

Find logic to perform multiplication of two 2-bit (unsigned) integers (i.e., 0–3), with a 4-bit output. *Hint*: use a separate Karnaugh map for each output bit.[1]

Problem 10.15

Find logic to perform addition of two 2-bit (unsigned) integers (i.e., 0–3), with a 3-bit output. *Hint*: use a separate Karnaugh map for each output bit.

Problem 10.16

(a) Write down the Karnaugh map for the function of the Boolean variables A, B, C, and D, which is true when $ABCD$ (when converted to decimal as an unsigned integer, D being the LSB) is odd *and* greater than 4.

(b) Find a (reasonably simple) boolean expression for the logic function you diagrammed in (a).

(c) Sketch a logic implementation of this function in terms of only 2-input `NAND` gates.

[1] Paul Horowitz and Winfield Hill, *The Art of Electronics*, 2nd ed. (Cambridge, 1989), Exercise 8.14 (ISBN: 0521370957).

Chapter 11

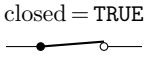
Physical Implementation of Logic Gates

So far, we have studied logic and logic gates, but logic is much more useful if we can implement logic gates *physically*. Generally speaking you can buy these as prepackaged integrated circuits, but it is still useful to understand how to implement these, for (1) extra intuition and (2) to understand the limits and quirks of commonly available electronic logic gates. We will start with simple examples of logic realizations and progress to realistic (but more complicated) cases.

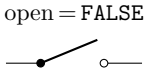
The material here in this chapter relies on previous material on diodes from Chapter 3 and transistors from Chapter 4. However, we will *briefly* review some of the relevant material here.

11.1 Simple Mechanical Switches

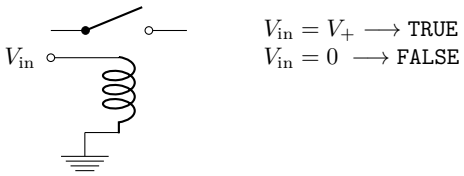
Fundamentally, electronic logic gates work by involving switches. Typically these are some form of *electronic* switches, but of course these can be ordinary mechanical switches (equivalent to connecting two points by a wire or breaking the wire connection). One simple convention for a **single-pole, single-throw (SPST)** switch ("single pole" = single circuit to break or connect; "single throw" = single possible connection to make or break), as shown below, is that the closed (conducting) or `ON` state is `TRUE`,



and the open or `OFF` state is `FALSE`.

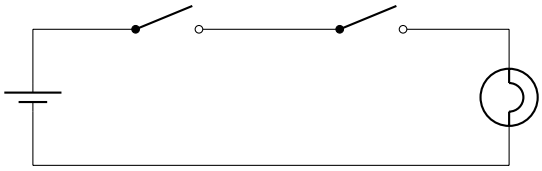


This convention agrees with a common convention for logic in terms of voltage levels, where `HIGH` voltage is `TRUE` and `LOW` voltage is `FALSE`, if we consider a **relay** (magnetically controlled switch), as shown below.

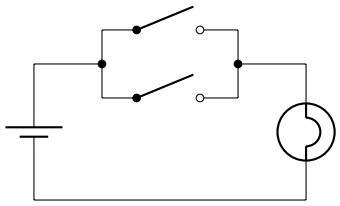


The relay is pulled closed when the voltage is `HIGH` (at some voltage V_+), due to the magnetic field of the coil; when the voltage is zero, there is no field and the switch pops open (due to the action of a spring).

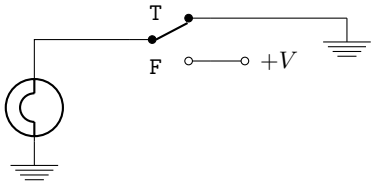
Using switches it is easy to see how to construct an `AND` gate, if two switches are in series, since both switches must close to light the light bulb (the logical "output" here).



For an `OR` gate, the two switches are in parallel, so only one switch needs to be closed to light the bulb.



As a final example, an inverter is shown below.



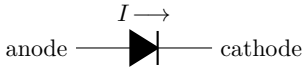
The switch here is a **single-pole, double-throw (SPDT)** switch ("double throw" = two alternative contacts for the switch), where "up" on the switch is `TRUE` and "down" on the switch is `FALSE`.

11.2 Diode Logic (DL)

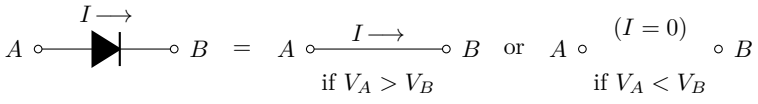
The simples "purely electronic" examples of logic come in the form of **diode logic (DL)**. Before examining some DL gates, first let's review how diodes work.

11.2.1 Diode Review

A diode is a two-terminal device, as shown below, and it acts as a one-way valve for current: current can only flow from the anode to the cathode (in the direction of the "diode arrow" in the schematic symbol).



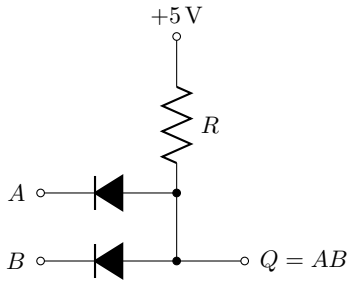
That is, as in the diagram below, if the anode voltage V_A is greater than the cathode voltage V_B, then current flows; otherwise, no current flows.



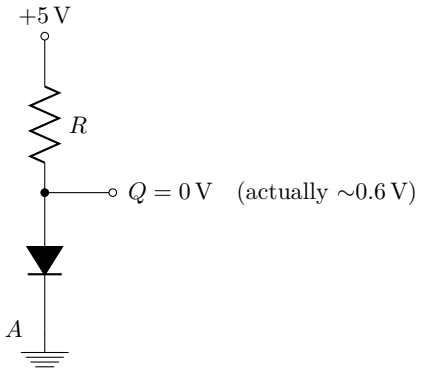
You can think of the diode as being a short circuit in the first case ("forward-biased"), and an open circuit in the second ("reverse-biased"). However, the real situation is a bit more complicated: a slightly better model is that there is a forward voltage drop of around 0.6 to 0.7 V when the diode is conducting current.

11.2.2 DL AND Gate

Now to see how to realize gates in DL. Below is a realization of an `AND` gate. The DL convention here is that 0 V is `FALSE`, and +5 V is `TRUE`.



If both inputs are at +5 V then all points are at the same voltage, including the output Q. If one input is low, say A, then the situation is as shown below.

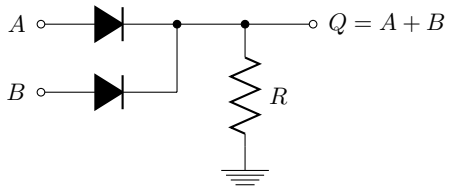


The diode is forward-biased, and thus shorts the output Q to ground. The power-supply voltage (+5 V) is dropped across the resistor because the diode causes sufficient current to flow through the resistor to ground to ensure this. The state of the other input (B) is irrelevant here, because either it "agrees" with A, or if it is `HIGH`, B's diode is reverse-biased, so it is disconnected from the circuit.

Actually, the output voltage is not quite 0 V in the latter case; because the diode has a forward-voltage drop, the output `FALSE` state is more like 0.6 V.

11.2.3 DL OR Gate

Another DL gate, the `OR` gate, is shown below.



Here, if either input is at +5 V, then the corresponding diode is forward-biased, so output is at +5 V [actually, (+5 − 0.6) V if we account for the diode's voltage drop]. If both inputs are at 0 V, then the whole circuit, including the output, is also at 0 V.

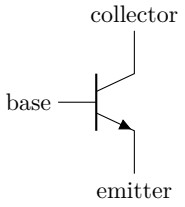
The main problem in the DL circuits is that one of the signal-voltage states "degrades" by 0.6 V on each gate, so not many gates can be cascaded while keeping the signal levels distinguishable. This motivates the use of *active* devices in logic circuits that can maintain the proper voltage levels.

11.3 Resistor-Transistor Logic (RTL)

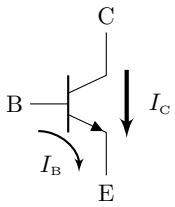
A step up in terms of sophistication is **resistor-transistor logic (RTL)**, which is obsolete but relatively easy to understand. Again, we first have to review how a transistor—specifically, the NPN bipolar junction transistor (BJT)—behaves, in particular as a switch.

11.3.1 BJT Review

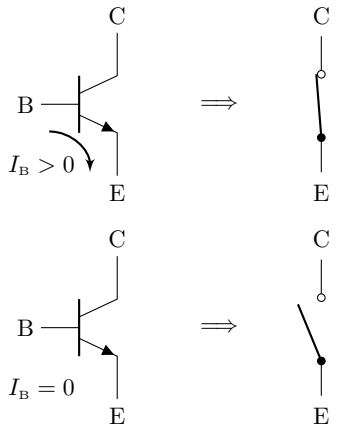
Recall that a transistor is a three-terminal device, with terminals labeled as in the diagram below.



The transistor acts as a switch for current, based on another current. We will consider two currents, I_B from the base to the emitter, and I_C from the collector to the emitter, as shown below.



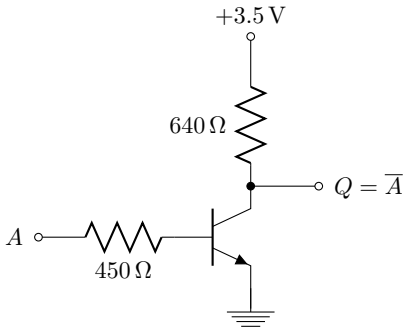
Then I_B acts as the control current, and I_C is the current to be switched. Simplistically, if there is some current I_B, then I_C can flow, so the C–E path acts as a closed switch.



However, if $I_B = 0$, then the C–E path acts as an open switch. There are some extra voltage drops to consider here, but this simple model suffices to understand RTL-gate operation.

11.3.2 RTL NOT Gate

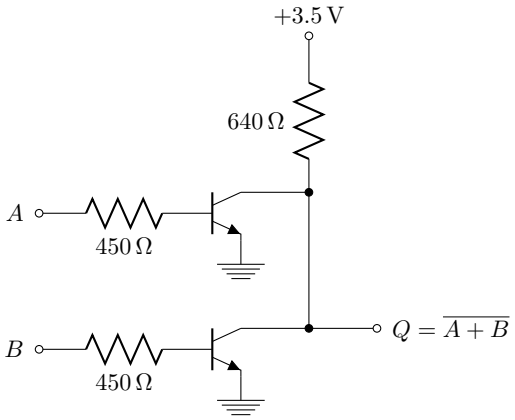
The RTL convention is that +3.5 V is `TRUE`, with 0 V `FALSE`. The simplest RTL gate is an inverter or `NOT` gate, shown below.



If the input is `TRUE`, then $I_B > 0$, and the C–E path conducts. This pulls the output down near ground, or `FALSE`. If the input is `FALSE`, then $I_B = 0$, and the C–E path is broken. The resistor pulls the output up to the supply voltage, or `TRUE`.

11.3.3 RTL NOR Gate

A slightly more complicated example is the `NOR` gate, shown below.

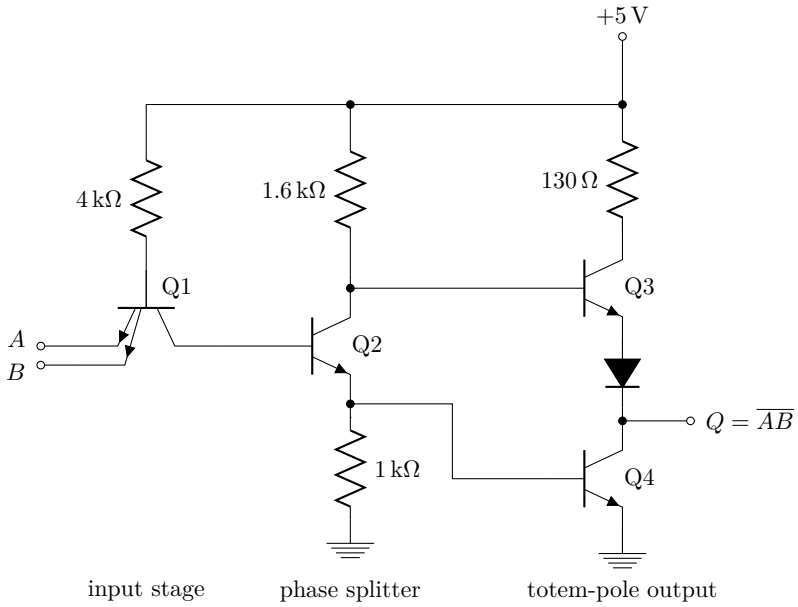


The operation is the same as the `NOT` gate, but here either input can pull the output to ground; the output is only pulled up `HIGH` in voltage if both inputs are `FALSE` or `LOW`.

RTL works reasonable well and doesn't suffer from the (cumulative) degradation problems of DL, because the output levels are set by the supply levels, not the inputs. However, DL is obsolete because the "return" to the high state when the transistors stop conducting is via the pull-up resistor. This transition is slow if there is a significant capacitive load on the output. (The `LOW` transitions when the transistors conduct are fast because the BJT collectors have effectively a very low impedance.)

11.4 The Real Thing: Transistor-Transistor Logic (TTL)

A common standard still in modern use is **transistor–transistor logic (TTL)**. The nominal convention is that +5 V is `TRUE`, and 0 V is `FALSE`. The circuitry is somewhat more complicated, and we'll go through the classic TTL `NAND` gate, shown below, as an example.



There are three different stages: the input stage (Q1), the phase splitter (Q2), and the totem-pole output (Q3, Q4, and the diode). An unusual feature is the double-emitter input transistor Q1. It works just like a regular transistor, except that a base current to *either* emitter will switch the collector current. Let's trace the voltages through the circuit for two cases.

1. Suppose A or B is `LOW`. Then:
   - Q1 is `ON` (collector conducts to grounded input).
   - Q2's base is `LOW`, thus Q2 is `OFF`.
   - Q3's base is `HIGH` (pulled up by 1.6-kΩ resistor), thus Q3 is `ON`.
   - Q4's base is `LOW` (pulled down by 1-kΩ resistor), thus Q4 is `OFF`.
   - The output is `HIGH` since it is pulled up via Q3 and the diode. The output is 5 V − Q3's voltage drop − the diode drop, which works out to around 3.5 V.
2. Suppose A and B are both `HIGH`. Then:
   - Q1 is `OFF` (collector disconnected from inputs).
   - Q2's base is `HIGH` (pulled up via the B–C path of Q1, which acts like a diode), thus Q2 is `ON`.
   - Q2's emitter is pulled `LOW` by Q4, which is `ON`.
   - Q2's collector is pulled `LOW` since it is `ON`; so Q4's base is `LOW`, and Q3 is `OFF`.
   - The output is `LOW` since it is pulled down via Q4. The output is 0 V + Q4's voltage drop, which works out to around 0.4 V.

The point of all this is to generate a few useful and general observations.

1. The inputs "want" to be high, because they tend to be pulled up to the power-supply voltage via the 4-kΩ resistor and the B–E paths of Q1. Thus, the inputs *source* current when they are pulled `LOW`. In particular, open inputs are `HIGH` by default in TTL, and less current flows (less power is dissipated) when the inputs are `HIGH`. In particular, if you have unused inputs in TTL circuits, it is best to tie them `HIGH` (i.e., connect them to +5 V).
2. The output, when driving another TTL input, must *sink* current, roughly $(5\,\text{V})/(4\,\text{k}\Omega) = 1.25\,\text{mA}$. One output can drive multiple inputs, but there is a limit to this, because the output has a limited current capacity. This limit is called **fanout**, which is typically ∼10 inputs for a standard TTL output.

3. The output voltages don't quite match the nominal values of 0 V and +5 V, so the TTL standard defines precisely the tolerance limits on signal voltages.
   - TTL circuits must recognize anything from +2.0 V to +5 *V* as `HIGH`.
   - TTL circuits must recognize anything from 0 V to +0.8 *V* as `LOW`.
   - The intermediate range of +0.8 V to +2.0 *V* is *indeterminate*: TTL circuits can do anything with inputs in this range and still conform to the TTL standard.

11.4.1 TTL Nomenclature

Standard TTL chips are most famously grouped into the 74XX (or 74XXX) family. For example, there are:

- 7400: quad, 2-input `NAND` (i.e., 4 `NAND`'s per package)
- 7402: quad, 2-input `NOR`
- 7404: hex inverter (i.e., 6 `NOT` gates)
- 7408: quad, 2-input `AND`

and there are hundreds more, though many are now becoming obsolete. Note that these are also labeled as equivalent 54XX circuits, which are the military-grade versions.

These "classic" TTL circuits are now obsolete, but they still come in many popular "flavors." These variations are labeled by a tag between the 74 and XX, for example 74LS00, 74F00, and 74HCT00 are all basically the same as the original 7400. The common flavors are:

- L: low-power (slow, obsolete)
- H: high-speed (high-speed, obsolete)
- S: high-speed Schottky (high-power, obsolete)
- LS: low-power Schottky (common, modern-standard chip)
- AS, ALS: "advanced" S, LS
- F: fast (gates have ∼4-ns propagation delay vs. ∼10 ns for standard gates)

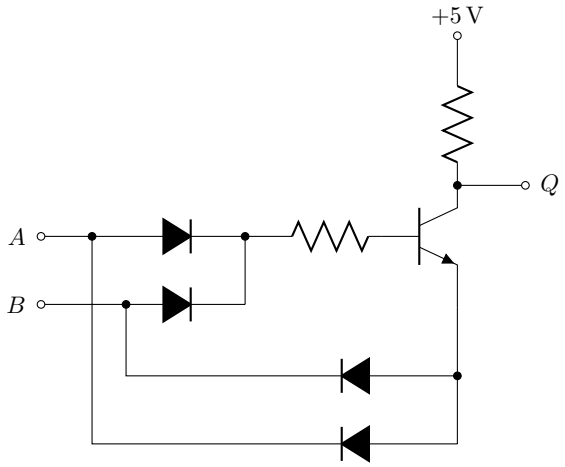
11.4.2 CMOS

Another class of devices is **complementary MOSFET (CMOS)complementary MOSFET**. These devices are similar to BJT logic devices, but are switched by voltage, rather than current (remember no current flows into the gate of a MOSFET). A side effect of CMOS designs is that they dissipate current while switching states, but not when "holding." This feature makes CMOS generally power-efficient compared to TTL, and CMOS circuits take negligible steady-state input current (but of course are more susceptible to static discharge). CMOS devices are more flexible in terms of logic level, and can operate at `HIGH` voltages other than +5 V. There are separate CMOS logic families, but there are also CMOS variations of standard 74XX devices. For example, the flavors are:

- C: (e.g., 74CXX) operates from +3 to +15 V (compared to +5 V for standard TTL), with a nominal "trigger" point of ½ of the supply voltage.
- HC: high-speed CMOS
- HCT: high-speed, compatible levels with TTL (+5 V, and a low trigger voltage)
- AC, ACT: advanced CMOS (i.e., fast), ACT is the advanced HCT

11.5 Circuit Practice

As circuit practice, consider the circuit below. What kind of gate is this?



Solution. Due to the arrangement of diodes, the only way to turn the transistor `ON` is to have one input `HIGH` and one `LOW`. In this case, the output is `LOW`; otherwise the output is `HIGH`. This, this is an `XNOR` gate.

11.6 Exercises

Problem 11.1

You have two switches (two-position switches; review how switches work if you need to!), a battery, a light bulb, and an infinite supply of wire. Devise a way to realize an `XOR` gate, where the switch positions are the inputs and the light bulb is the "output." How about an `XNOR` gate?

Problem 11.2

You have three switches (two-position switches; review how switches work if you need to!), a battery, a light bulb, and an infinite supply of wire. Devise a way to realize the logic expression $A \cdot (B + C)$, where the switch positions are the inputs and the light bulb is the "output."

Chapter 12

Multiplexers and Demultiplexers

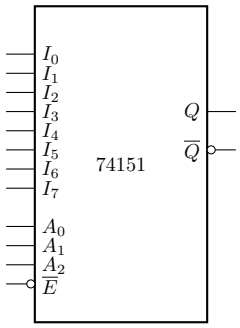
12.1 Multiplexers

Simply put, a **digital multiplexer** (or **MUX** for short) is a logic device that maps one of many (digital) inputs to one (digital) output. You select which input to connect to the output using the "address" inputs. The multiplexer is the logic analog of a many-to-one mechanical rotary switch.

Multiplexers are useful devices. For example, you can use them to "pack" data from multiple sources ("parallel data") onto a single "serial" transmission line (e.g., for phone or computer networks). They can also be used to sample or "poll" data from multiple sources, and ultimately allow scaling of many digital devices into modern computers. Multiplexers are examples of **MSI (medium-scale integration)** devices, "medium-scale" here meaning dozens of gates on 1 chip.

12.1.1 Example: 74151

An example of a multiplexer is the 74151 (which for example, with manufacturer and TTL-flavor codes would be something more like DM74LS151), an 8-input MUX, shown schematically below.



There are a number of features here:

- I_0–I_7 are the 8 inputs.
- A_0–A_2 are the 3 address lines, to select among the $2^3 = 8$ inputs; the idea is to select input n by setting $A_2A_1A_0$ to n in binary.
- Q is the output: the selected input is copied to the output.
- $\overline{Q}$ is an inverted copy of the output.
- $\overline{E}$ or $\overline{\texttt{ENABLE}}$ (also called "`STROBE`") is a "chip enable" line. If $\overline{E}$ is `LOW`, the chip works as we have described; if $\overline{E}$ is `HIGH`, then $Q =$ `LOW` $\overline{Q} =$ `HIGH`, independent of the states of I_{0-7} and A_{0-2}.

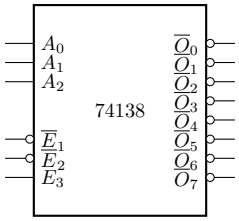
Another example of a common MUX is the 74150, a 16-input MUX (with 4-bit address).

12.2 Demultiplexers

A **demultiplexer** (or **DEMUX** for short) is the "opposite" of the MUX, in the sense that a single input is copied to a selected one of many possible outputs. Again, these are useful in, for example, packing and unpacking data to and from a transmission line via a MUX–DEMUX pair. Also a variation on the DEMUX is a **decoder**, which is the same as a DEMUX, but only *selects* the output, without copying any input (the "data" is effectively constant `HIGH`).

12.2.1 Example: 74138

A good DEMUX example is the 74138, a 1-to-8 DEMUX, as shown below.



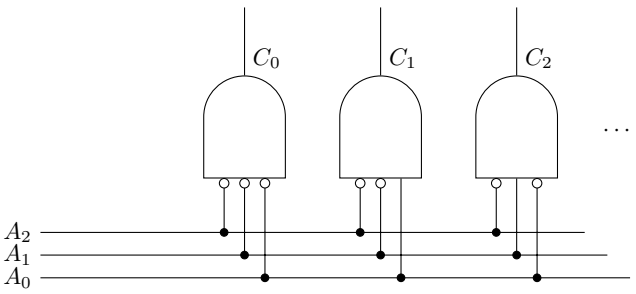
To go over the features here:

- $\overline{O}_0$–$\overline{O}_7$ are the 8 outputs, Note that they are inverted (i.e., their "normal," unselected state is `HIGH`).
- A_0–A_2 are the 3 address lines, again to select among the $2^3 = 8$ inputs in the same way as the MUX.
- $\overline{E}_1$, $\overline{E}_2$, and and E_3 are chip-enable inputs. The chip is enabled if $\overline{E}_1 = \overline{E}_2 =$ `LOW` and $E_3 =$ `HIGH`. Then the operation is as follows.
  - If the chip is enabled, then the selected output $\overline{O}_j$ is `LOW`. (The others are `HIGH`.) In this case, the chip acts as a decoder.
  - If the chip is *not* enabled, then all outputs $\overline{O}_j$ are `HIGH`.
  - To operate this chip as a DEMUX instead of just a decoder, use $\overline{E}_1$ *or* $\overline{E}_2$ as a data input. In this case, the selected output copies $\overline{E}_1$ or $\overline{E}_2$, while the others remain `HIGH`. Alternately, E_3 can work as a data input, in which case the selected output copies $\overline{E}_3$ (with the others still `HIGH`).

Another example of a common MUX is the 74154, a 16-output decoder/DEMUX (with 4-bit address).

12.3 Making a MUX

The logic underlying a multiplexer is not difficult to understand. There are two basic elements: a decoder and "routing" logic. As an example, let's consider an 8-input multiplexer. The decoder, as we described, takes address inputs A_0–A_2, and sets the corresponding one of 8 outputs C_0–C_7 `HIGH`, with the others `LOW`. We can simply use `AND` gates to set each output when matching the correct address combination, as below.



To be more efficient in terms of gates, we would only use one NOT gate for each of $\overline{A}_0$, $\overline{A}_1$, and $\overline{A}_2$, rather than NOT gates as shown, but the input NOT operations simplify the diagram.

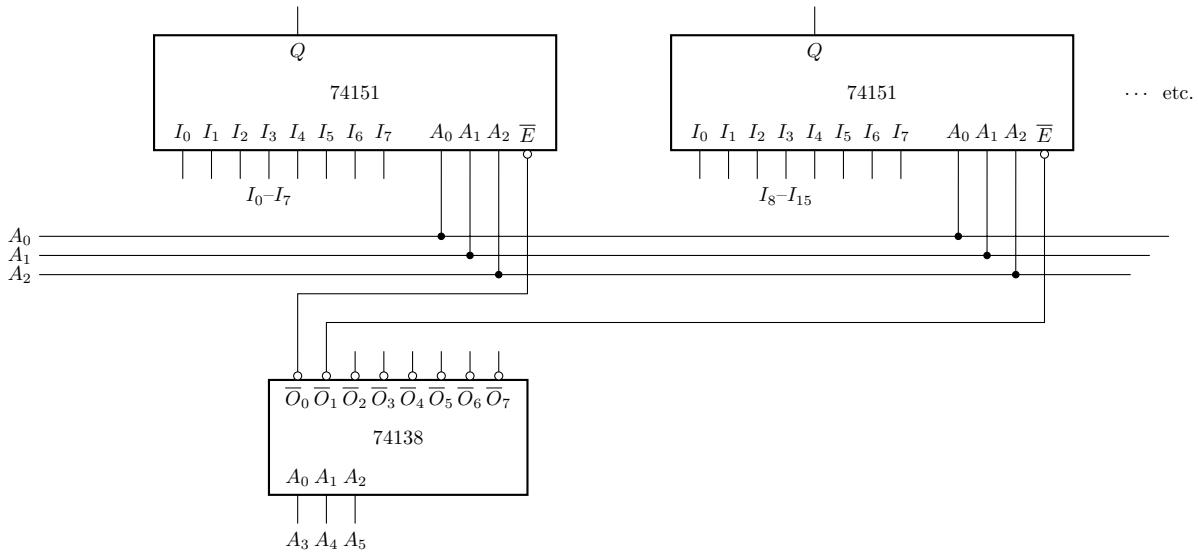
The routing logic then has the algebraic form

$$Q = (C_0 I_0 + C_1 I_2 + \cdots)\overline{E}, \tag{12.1}$$

if we include an $\overline{\texttt{ENABLE}}$ input.

12.4 Expanding a MUX (or DEMUX)

A useful technique is to combine MUXs into larger MUXs. For example given, 8× 8-input MUXs, how do we make a 64-input MUX? This shows the real idea behind having chip-enable inputs: we will use the $\overline{E}$ inputs and an 8-output decoder, as shown below.



One detail that we have left out is that the Q outputs must be combined by an OR gate. An important alternative is to use chips with **three-state logiclogic!three-state**. For these chips, the output is *disconnected* (high-impedance) when the chip is not enabled. In this case, you can just connect the chip outputs directly together, since only one chip will be enabled at a time. In this example, we can use the three-state alternative 74251 instead of the 74151.

12.5 Analog MUX/DEMUX

The same ideas behind *digital* MUX/DEMUX can apply to analog signals as well. An **analog switch** (or **CMOS switch**) is an electronic switch for analog signals, controlled by a digital input. An example is the DG412 quad SPST (normally open) analog switch. The switches are switchable electronically, e.g., from a computer-interface output. They are even good compared to mechanical switches in terms of noise: for example, you can run a digital control wire to the front panel from a circuit board rather than a signal line, and the noise pickup is not critical for a digital control line (it is easier to keep low-noise, critical signals on a well-grounded circuit board than to carry the signals on wires away from the board).

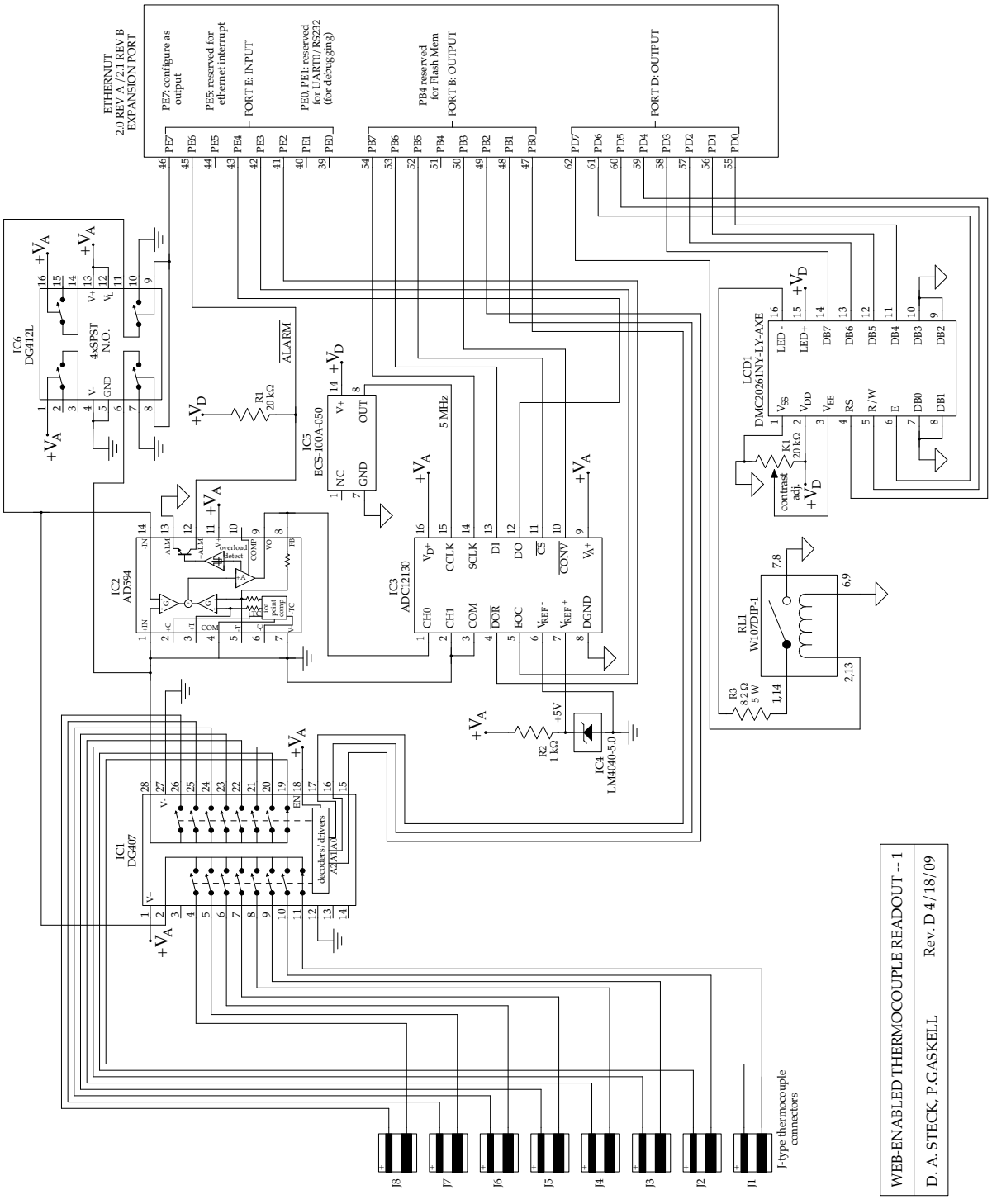
An analog MUX/DEMUX (in the analog case, these devices act as bidirectional devices, so there is no distinction between MUX and DEMUX) is an array of analog switches, controlled by address lines. This is really the analogue of a mechanical rotary-select switch. A good application here, for example, is to connect many sensors to a single microcontroller. An example of an analog MUX/DEMUX is the DG407, a dual 8-channel MUX.

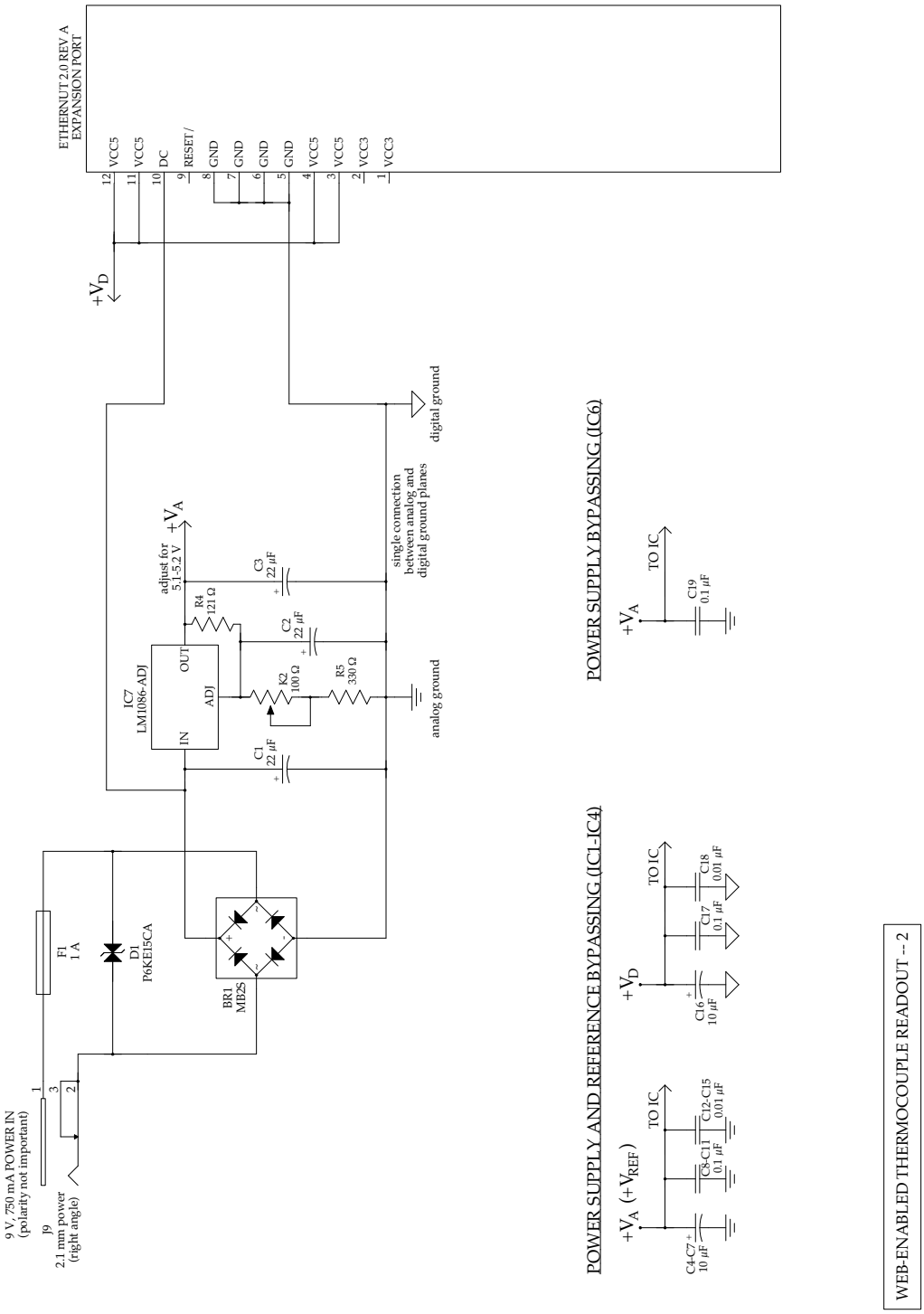
12.6 Circuit Practice: Multiplexed Thermocouple Monitor

On the following pages, look over the schematic for a web-enabled thermocouple monitor.[1] A few things to look over:

- There are provisions for 8 thermocouple inputs.
- The thermocouples are monitored by the AD594, which provides "ice point" compensation and buffers the thermocouple signal. However, this is a relatively expensive chip, so rather than having one for each thermocouple, we just use an analog multiplexer (IC1).
- The output is converted to digital via an analog-to-digital converter (IC3).
- The multiplexer address is controlled by an Ethernut microcontroller, which also reads out the ADC. The Ethernut has an ethernet port, and thus has firmware to make a web site to display the temperatures. It also controls an LCD display on the actual box.

[1] `http://atomoptics-nas.uoregon.edu/~zoinks/#WebTC`


ETHERNUT 2.0 REV A / 2.1 REV B EXPANSION PORT
PE7: configure as output
PE5: reserved for ethernet interrupt
PORT E: INPUT
PE0, PE1: reserved for UART0 /RS232 (for debug,prog)
PB4 reserved for Flash Mem
PORT B: OUTPUT
PORT D: OUTPUT
IC6 DG412L
4xSPST N.O.
IC5 ECS-I00A-050
5 MHz
ALARM
R1 20 kΩ
IC2 AD594
IC3 ADC12130
IC1 DG407
decoders/ drivers
IC4 LM4040-5.0
R2 1 kΩ
+5V
LCD1 DMC20261NY-LY-AXE
contrast adj.
K1 20 kΩ
RL1 W107DIP-1
R3 8.2 Ω 5 W
J-type thermocouple connectors
J8
J7
J6
J5
J4
J3
J2
J1
WEB-ENABLED THERMOCOUPLE READOUT -- 1
D. A. STECK, P.GASKELL Rev. D 4/18/09



WEB-ENABLED THERMOCOUPLE READOUT -- 2

12.7 Exercises

Problem 12.1

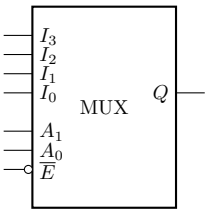
Show how to make a 4-input (digital) multiplexer from ordinary logic gates.[2]

Problem 12.2

Look up the 74139. What is it? Show how to hook it up, using at most an extra inverting buffer, to make an 8-output decoder. Show how (by adding extra logic gates) to implement an 8-output DEMUX.

Problem 12.3

Suppose that you end up stuck on an isolated desert island, surrounded by sand, coconut trees, and (oddly) a near-infinite supply of 4-bit MUXs, shown schematically below. Explain how you can, *in principle*, use your treasure cache to create *any* logic circuit you can dream up, to use to call for help. Make *specific* wiring diagrams to support your argument as necessary (you should have at least one diagram showing a wired-up MUX.) Assume you have found some MacGyverish way to adapt coconuts and coconut fibers to create whatever power supplies and wiring you need. However, you may *not* take the MUXs apart to obtain the individual logic gates inside.



[2]Paul Horowitz and Winfield Hill, *The Art of Electronics*, 2nd ed. (Cambridge, 1989), Exercise 8.17a (ISBN: 0521370957).

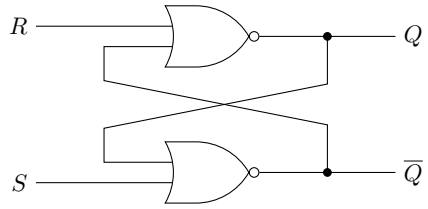
Chapter 13

Flip Flops

13.1 Flip-Flop Construction: SR Flip Flop

A **flip-flop** is relatively simple logic circuit that involves feedback (i.e., such that the output of a gate drives its own input, generally via other gates). Flip-flops are useful devices, and as we will see, they are the basis of digital memory.

The basic flip-flop is the **SR flip-flop** ("SR" for "set–reset"). A realization in terms of `NOR` gates is shown below.



To analyze this, let's work out the truth table.

S	R	Q	$\overline{Q}$
1	0	1	0
0	1	0	1
0	0	1	0
0	0	0	1
1	1	0	0

A few things to notice here: First, there are *two* rows with inputs $SR =$ `00`, with different outputs. You should convince yourself that both are consistent with the circuit. In the first two rows, the fact that one input is `HIGH` fixes the state of the corresponding `NOR` gate, which then fixes the state of the other one. But in this multivalued, or **bistable** state, the inputs don't fix the state of either gate. Rather, we have to *assume* that Q is in some state (i.e., it was set in this state in the past), which then fixes $\overline{Q}$. **This bistable state is the defining characteristic of a flip-flop**: it means there is **hysteresis** in the circuit, so that the state of the circuit "remembers" the past state. It is in this sense that a flip-flop can act as memory.

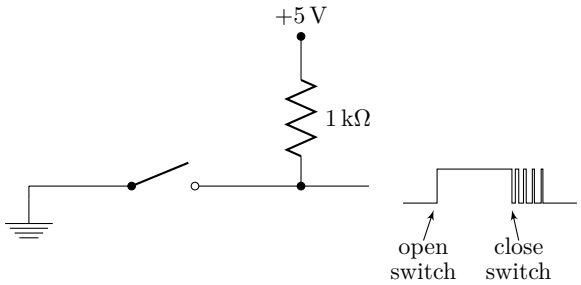
A second feature in the truth table is that the state $SR =$ `11` is a "bad" state, since $Q = \overline{Q}$, which means our output notation is in some sense itself bad. However, having the complementary outputs is convenient, even if nonessential. The more important problem with this state, however, is that the outputs don't match either of the two "hysteresis states," which we want to use as memory. So if we take the inputs from the bad state to $SR =$ `00`, it will collapse into one or the other hysteresis state in an ill-defined way, which is not very useful. Generally speaking, the bad state is to be avoided when using a flip-flop for its intended purpose.

Then this is how you use a flip-flop:

- The inputs R and S are normally `0` (i.e., the flip-flop is in one of the memory states).
- Bringing S to `1` and back to `0` (the "set" operation) changes Q = `1` and $\overline{Q}$ = `0`. This state is "remembered" when $R = S = 0$.
- Bringing R to `1` and back to `0` (the "reset" operation) changes Q = `0` and $\overline{Q}$ = `1`. This state is "remembered" when $R = S = 0$.

13.1.1 Application: Debounced Switch

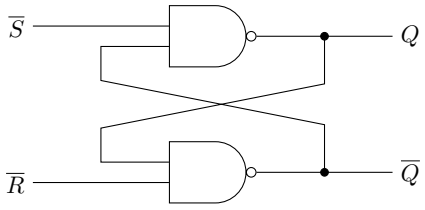
A simple application of the SR flip-flop is to make a **debounced switch**. Recall that switches are mechanical devices that make and break electrical connections. We can use a switch as in the schematic below to toggle between TTL `HIGH` and `LOW`.



That is, if the switch is open, the output is pulled up by the resistor to +5 V, while a closed switch corresponds to a 0-V output.

The problem is that the output will really look like the output shown for one open/close cycle. When we open the switch, the output goes `HIGH` with no problem, because the switch cleanly breaks the connection. However, when closing the switch, there is a problem. The contacts must close, and normally they are held together by some spring pressure. But when they close, one contact smacks into the other and "bounces" off of it, just like dropping a chunk of metal on a hard floor. The spring action pushes the contacts together again, and the result is a few extra, short pulses due to the switch bounce, typically on ms time scales.[1] This is a real problem, for example, if the pulse is to drive the input of a counter. For example, the switch could be actuated by items on a manufacturing line, to count the number of items produced; it would obviously not be a very good count if there were several extra bounces for each item to count.

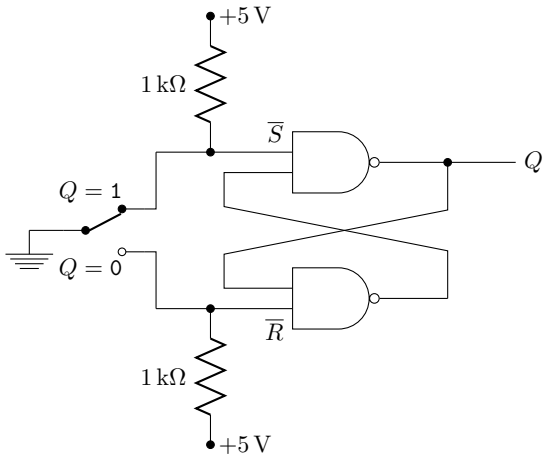
A simple solution to this uses a flip-flop and a slightly more complicated switch. Before getting to that, let's introduce a functionally equivalent variant of the above RS flip-flop, now based on `NAND` gates.



The operation is the same as before, but note the inputs are $\overline{R}$ and $\overline{S}$, so their senses are inverted. That is, the "usual" input state should be $\overline{R} = \overline{S}$ = `1`. Then you bring $\overline{S}$ momentarily to `0` to set the flip-flop (i.e., Q = `1`), and you bring $\overline{R}$ momentarily to `0` to reset it (Q = `0`). We will leave the analysis of this flip-flop as a circuit-practice exercise.

Now the debounced switch uses an SPDT switch (the "bouncy" switch used an SPST switch). The "up" switch state sets Q = `1`, and the "down" switch state sets Q = `0`.

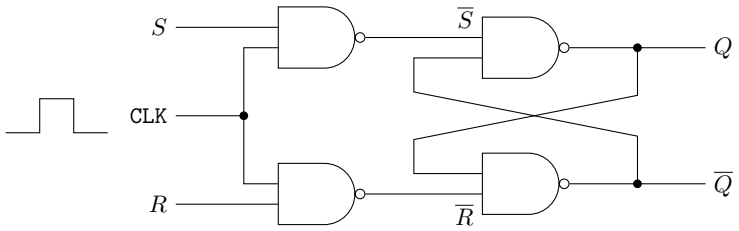
[1] example: `http://www.maximintegrated.com/en/app-notes/index.mvp/id/287`



During a bounce, the switch makes no connection to either contact, so both $\overline{S}$ and $\overline{R}$ are `1`. This is the memory state, so the flip-flop holds the last switch state, which persists through the duration of the bouncing.

13.2 Clocked Flip-Flops

An important class of flip-flops, one step up in sophistication from the basic flip-flops above, is that of **clocked flip-flops**. A **clock** is an external, typically periodic logic signal that synchronizes signals in complex circuits. We will see some examples later, but in complex circuits, this synchronization is important in avoiding problems with race conditions. The idea in a clocked flip-flop is that the input datum is only accepted during a particular phase of the clock cycle, for example when the clock is `HIGH`. The clock then functions as a "gate" for the input data. An example of a clocked SR flip-flop is shown below.



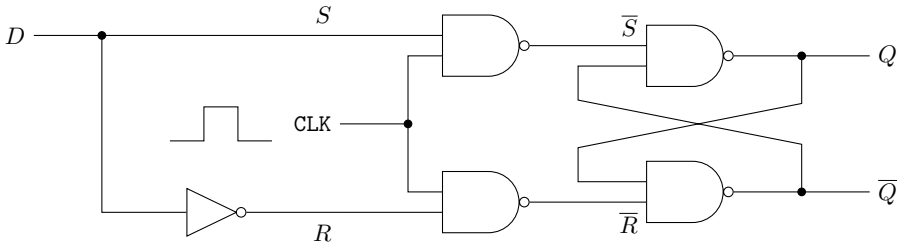
Note that when the `CLK` signal is `LOW`, this guarantees that the flip-flop is in the memory state. We can write the truth table for this circuit as follows.

S	R	Q_{n+1}
0	0	Q_n
1	0	1
0	1	0
1	1	"bad"

Here, Q_n is the output state after the nth pulse. If S and R are `LOW`, then the flip-flop stays in the memory state, and the state Q_n persists to the next clock cycle. Otherwise, the clocking `NAND` gates act as inverters for the S and R inputs, so the inputs set and reset as usual (with momentary `HIGH` action), but only when the clock is `HIGH`.

13.2.1 D-Type Flip-Flop

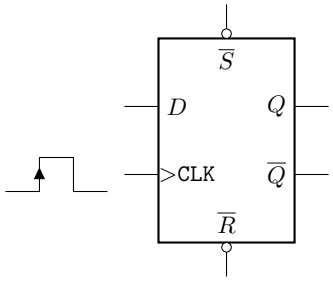
An important class of clocked flip-flop is the **D-type flip-flop**, which is basically the clocked SR flip-flop, but where the two inputs are always in opposite states, as shown below. Here, the D or "data" input drives the S and R inputs oppositely.



The idea is that there is now only one input, and the flip-flop latches the value of the data while `CLK` is `HIGH`. This flip-flop is often called a **data latch**.

13.2.2 Edge-Triggered, D-Type Flip-Flop

A somewhat more sophisticated and realistic D-type flip-flop is the **edge-triggered, D-type flip-flop**. Here, "realistic" means you can buy these prepackaged (e.g., the 7474 gives you two of these per chip). The main difference is that the data-latching action happens on the *rising edge* of the `CLK` pulse. Since the edge has a short duration compared to the `HIGH` phase of the clock, the timing is more precise in this convention. Schematically, this flip-flop is shown below.



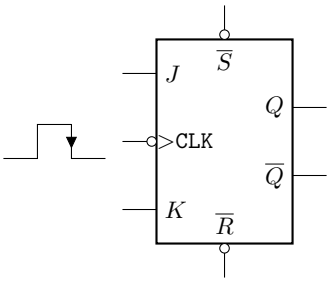
Some things to notice:

- D is the data input, as in the regular D-type flip-flop.
- `CLK` is the clock input. Again, the datum is latched on the rising edge. Schematically, it is common to indicate this by drawing an arrowhead on the edge of a sample clock pulse, as shown, and also to include a "$>$" next to the `CLK` pin.
- Q is the output as usual, and $\overline{Q}$ is an inverted output copy.
- $\overline{S}$ and $\overline{R}$ are "jam" set and reset inputs. These override the output, independent of the `CLK` state (so they work just like the inputs to the SR flip-flop). These are often called $\overline{\texttt{PRE}}$ and $\overline{\texttt{CLR}}$ (for preset/clear).

This flip-flop is good, for example, for storing data until they are "passed on" to a computer (e.g., in data-acquisition systems, when data arrive with timing determined by a physical system, but need to be loaded into a computer with its own timing).

13.2.3 JK Flip-Flop (Edge-Triggered)

A slightly more complicated variation on the edge-triggered, D flip-flop is the **edge-triggered, JK flip-flop**. This is like the D version, but there are *two* data inputs (J and K), with no indeterminate states for J and K. This is available in the 74112/74112A (2 per chip), and the now-obsolete 7476/7476A (also 2 per chip). The flip-flop is shown schematically below.

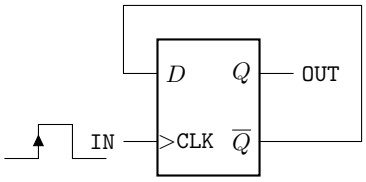


The operation with the two new inputs is as follows.

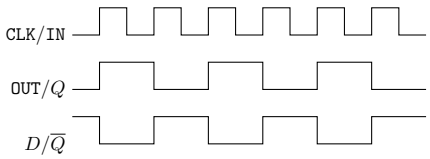
- The CLK on this flip-flop, as drawn, triggers on the *falling* edge of the clock. Note the NOT circle on the clock input, and the sample clock pulse. This is how the 74112/74112A and 7476A work (the plain 7476 triggers on the positive edge). That is, this device is **negative-edge triggered**.
- If $J =$ 0 and $K =$ 0, then Q persists on the next CLK pulse.
- If $J =$ 1 and $K =$ 0, then $Q =$ 1 on the next CLK pulse.
- If $J =$ 0 and $K =$ 1, then $Q =$ 0 on the next CLK pulse.
- If $J =$ 1 and $K =$ 1, then Q **inverts** on the next CLK pulse (i.e., it "toggles").
- $\overline{S}$ and $\overline{R}$ are still jam set and reset inputs.

13.3 Counters

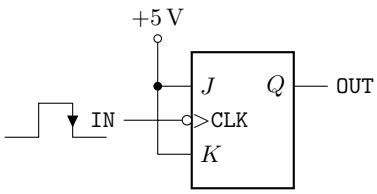
One useful application of flip-flops is in realizing **counters**, which count input pulses by incrementing a binary output. The basic building block of a counter is the **divide-by-2 counter**, shown below in terms of a D-type flip-flop.



The timing diagram for this circuit is shown below. Note that transitions happen on the rising edge of the input (clock) pulses, and essentially the output is just toggling its output on each clock cycle. Hence the term "divide-by-2," since the output pulse train oscillates at half the frequency of the input clock. More specifically, the flip-flop loads $D = \overline{Q}$ to Q on each rising pulse-edge.

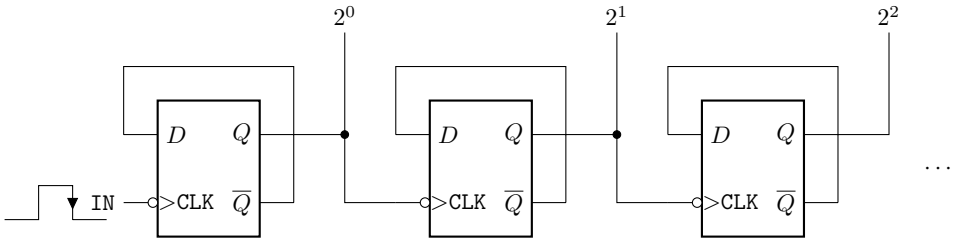


Since the output is just toggling, recall that we can also make a JK flip-flop do this by tying both J and K inputs HIGH, as shown below.

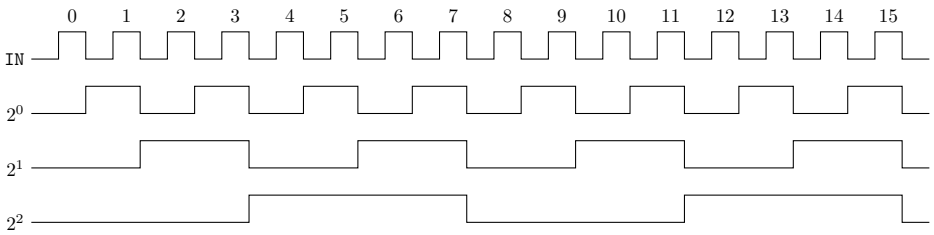


13.3.1 Asynchronous (Ripple) Counter

Generalizing the divide-by-2 is relatively easy. For example, we can make a divide-by-4 counter by cascading 2 divide-by-2 counters, and by chaining 3 of them, we make a divide-by-8 counter. Chaining n counters realizes a divide-by-2^n counter, as shown schematically for D-type flip-flops below (first three bits are shown).



The timing diagram is shown below. Note that we changed the convention for the flip-flops, which now trigger on the *falling* edge of the clock pulse. (Why do we need to trigger on the falling edge? How would you modify the circuit if the flip-flops triggered on a *positive* edge?)



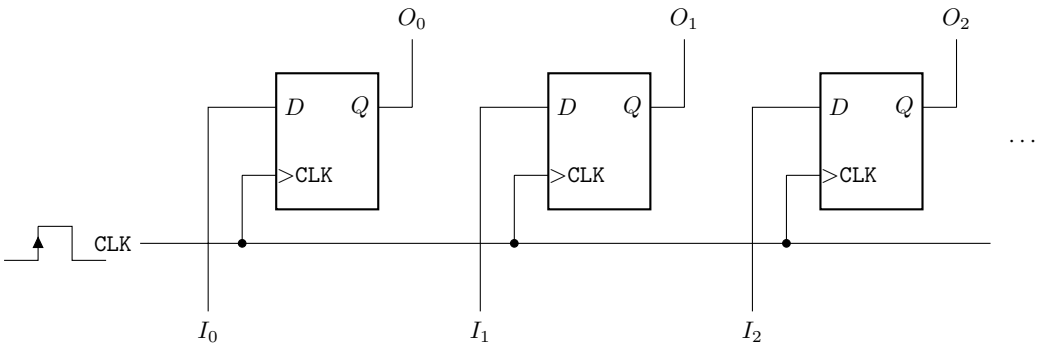
The main advantage of this circuit is that it is easy to build: it's easy to chain together flip-flops. The main disadvantage is the **asynchronous** or "**ripple**" operation of the circuit: since there is a finite propagation delay of the logic signal through each flip-flop, it takes some time for each clock pulse to "ripple" through a long chain of a many-bit counter, which can cause synchronization problems for fast input signals (i.e., spurious output states may be present for some or even all the time).

13.4 Memory and Registers

Flip-flops act as single-bit memory devices, as we have seen. Combining flip-flops, we can build up **registers**, which act as multi-bit memories.

13.4.1 Register

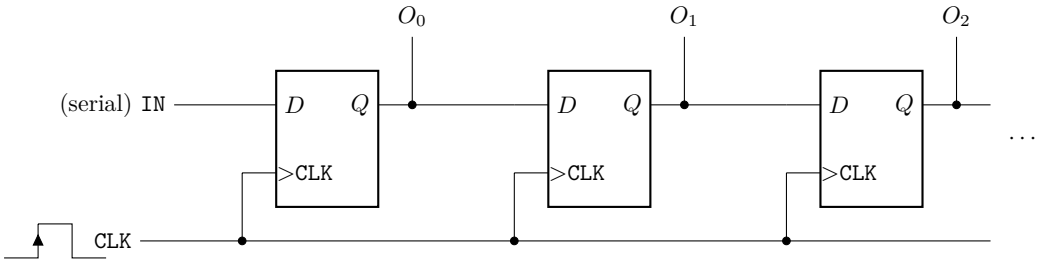
The basic register is an array of D-type flip-flops, which synchronized CLK inputs, the idea being to latch all the bits at once (to avoid timing problems, e.g., as in the ripple counter).



Once latched, the register holds the output state, independent of the inputs, until the next clock pulse. One application is where a shared set of data logic lines drives several devices; a register at the input of each device can hold the relevant logic data for each particular device while the data lines drive other devices.

13.4.2 Shift Register

The **shift register** shifts the data among the outputs, shifting all bits in one direction on each clock cycle.

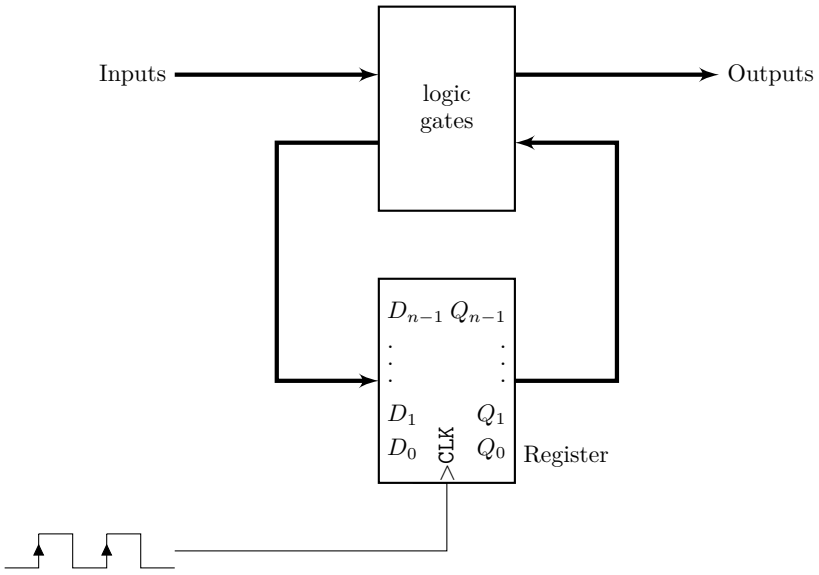


An example application is in converting serial data to parallel form (i.e., on the receiving end of a serial transmission channel). Also, note that a bit shift corresponds to a mathematical operation on binary data (divide/multiply by 2).

13.5 Sequential Logic and the State Machine

Recall in our discussion of asynchronous circuits (e.g., the ripple counter), we mentioned that there can be timing problems if the signals change rapidly, such that the gate delays are comparable to the time between transitions. The cure for this is to use **synchronous circuits**, where all logic transitions happen just after each clock pulse (or, more commonly and precisely, at an edge of each clock pulse). The transitions occur based on the logic levels present just *before* each clock pulse (edge). This is essentially what happens in a microprocessor, and this system of synchronous, clocked transitions is essential for high-speed and high-complexity logic systems.

The general scheme of sequential logic is shown in the diagram below.



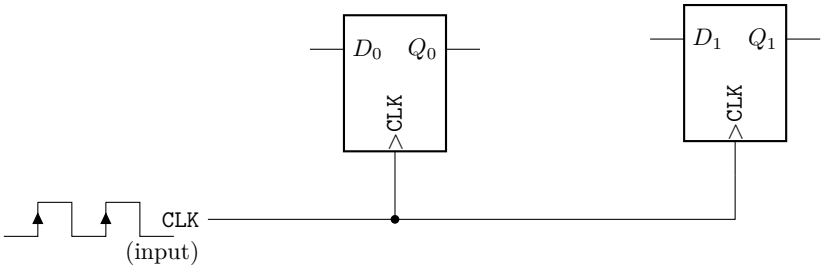
The first main ingredient is a register (Section 13.4.1), which is again an array of D-type flip-flops, with a common clock input. On the rising clock edge, the inputs D_j are all transferred to the outputs Q_j, and then held.

The next main ingredient is a set of logic gates in some combination, which work to transform the outputs Q_j into new inputs D_j. These can be implemented generically using **PAL (programmable array logic)** devices or **PLA (programmable logic array)** devices, which are basically configurable arrays of many logic gates. Also available are **registered PAL/PLA** chips, which contain flip-flops and gates all on one chip. These are usually called **PLD's (programmable logic devices)**.

Inputs and outputs to the logic gates also permit interaction with the outside world. The sequential logic scheme here is the most general form of digital logic, even though the idea is schematically simple.

13.5.1 Example: Synchronous, Divide-by-3 Counter

Continuing to follow Horowitz and Hill,[2] we will illustrate sequential logic by constructing a synchronous, divide-by-3 counter. We will need two flip-flops (two bits to accommodate counting to 3), clocked from the counter input. We will call the register inputs D_0 and D_1 with corresponding outputs Q_0 and Q_1.



To design the counter, first we will choose the sequence of states that we want. There are no external inputs here, so this is just a simple sequence, with no conditions (which we would represent as extra bits here). The counting sequence is thus as follows:

Q_0	Q_1
0	0
0	1
1	0
0	0

Here, Q_0 functions (arbitrarily) as the MSB, and Q_1 the LSB. We have also shown the first step in the repeat.

The next step is to find the appropriate D's. Remember the D's must be our desired Q's on the next step, so we will explicitly write out the desired D's as a function of the Q's.

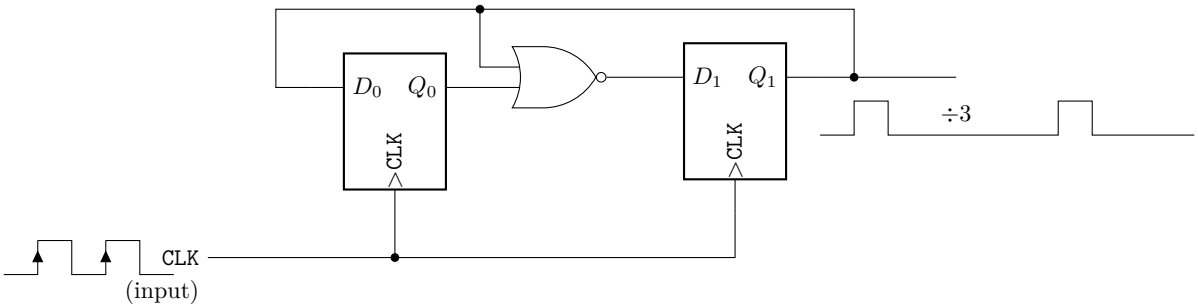
Q_0	Q_1	D_0	D_1
0	0	0	1
0	1	1	0
1	0	0	0
0	0	0	1

Finally we find a logic implementation of the functions $D_j(Q_k)$, using whatever means necessary (e.g., Karnaugh maps). Here, by inspection we can see that

$$D_0 = Q_1, \qquad D_1 = \overline{Q_0 + Q_1}. \tag{13.1}$$

Thus, the circuit implementing the counter is shown below.

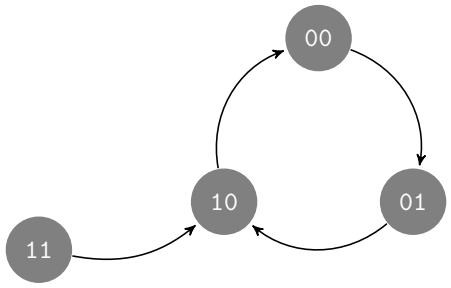
[2] Paul Horowitz and Winfield Hill, *The Art of Electronics*, 2nd ed. (Cambridge, 1989), p. 513 (ISBN: 0521370957).



One more detail to worry about is the set of **excluded states**. For the divide-by-3 counter, we didn't use the state $Q_0Q_1 = \texttt{11}$, but what if the flip-flops end up in this state (e.g., when the circuit is turned on)? Given our logic realization, we will then have $D_0 = \texttt{1}$ and $D_1 = \texttt{0}$, so the counter will resume the normal counting cycle on the next cycle, with $Q_0Q_1 = \texttt{10}$. But it's important to consider these, since the register could end up "frozen" in an excluded state or a cycle of excluded states if the logic gates don't handle them properly.

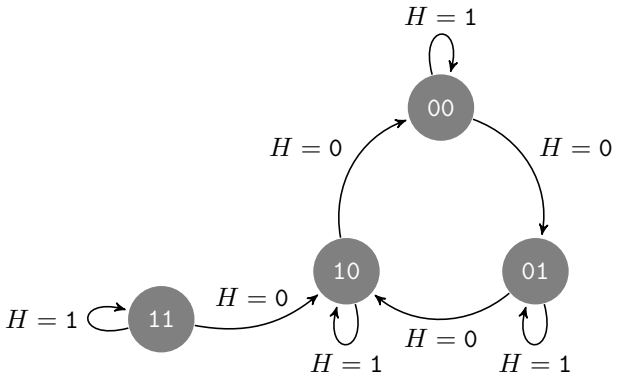
13.5.2 State Diagrams

A convenient way to represent the operation of a state machine is a **state diagram**. For example, the diagram (including the excluded state) for the divide-by-3 counter is shown below.



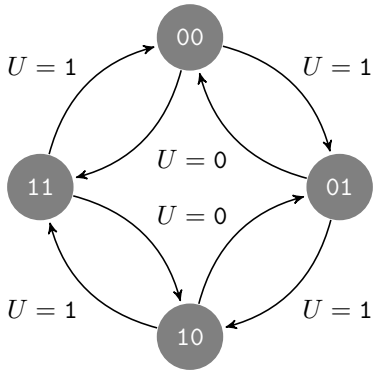
This is a directed graph of all the register states; the edges (arrows) show transitions between the states. You can use a diagram to start the design: you just begin with n flip-flops, where 2^n equals or exceeds the number of distinct states. Then you use the procedure that we used for the divide-by-3 counter to generate the appropriate logic.

If there are inputs, then there can be multiple possible transitions depending on the inputs. In this case, you can modify the diagram, labeling the transitions according to the input state. For example, below is a divide-by-3 counter with hold. That is, a "hold" bit H causes the counter to hold its state when $H = \texttt{1}$ and counts as before when $H = \texttt{0}$.



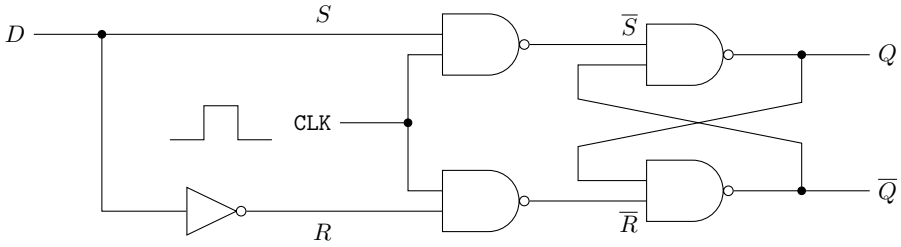
This example also illustrates "transitions" where the state remains the same. We will cover the implementation of this counter in Section 13.7.1.

Another example is the 2-bit up/down counter, where an input bit U controls whether the (divide-by-4) counter counts up or down.

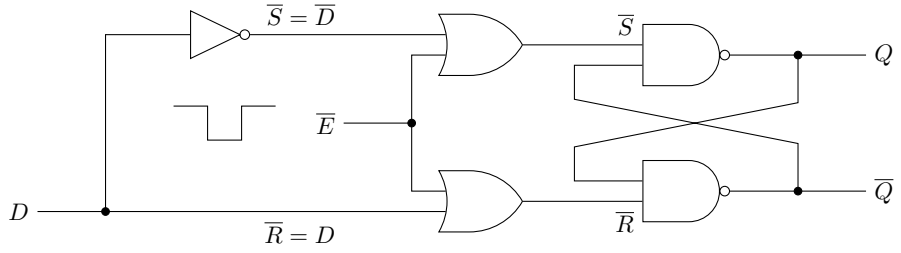


13.6 Memory

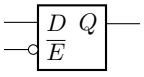
When we introduced the flip-flop, we noted that it is the basic building block of memory, since it is a simple 1-bit memory device. Now we will talk about how to build up many flip flops into a memory device. First, however, we will refer to the (non-edge-triggered) D-type flip-flop from Section 13.2.1. Recall that this is as shown below.



An RS flip-flop (with momentary-low inputs) has two extra gates on the inputs and a clock signal, so that the flip-flop only accepts input from the data line D when CLK = HIGH. The clocked version of the flip-flop is essential in scalable memories. We will actually refer to a modified version of this circuit, shown below.



Here we changed notation so that the clock signal is now an "enable-low" ($\overline{E}$), and the flip-flop now accepts input only when $\overline{E}$ = LOW. To keep things compact, we will refer to this circuit with the small block diagram below.



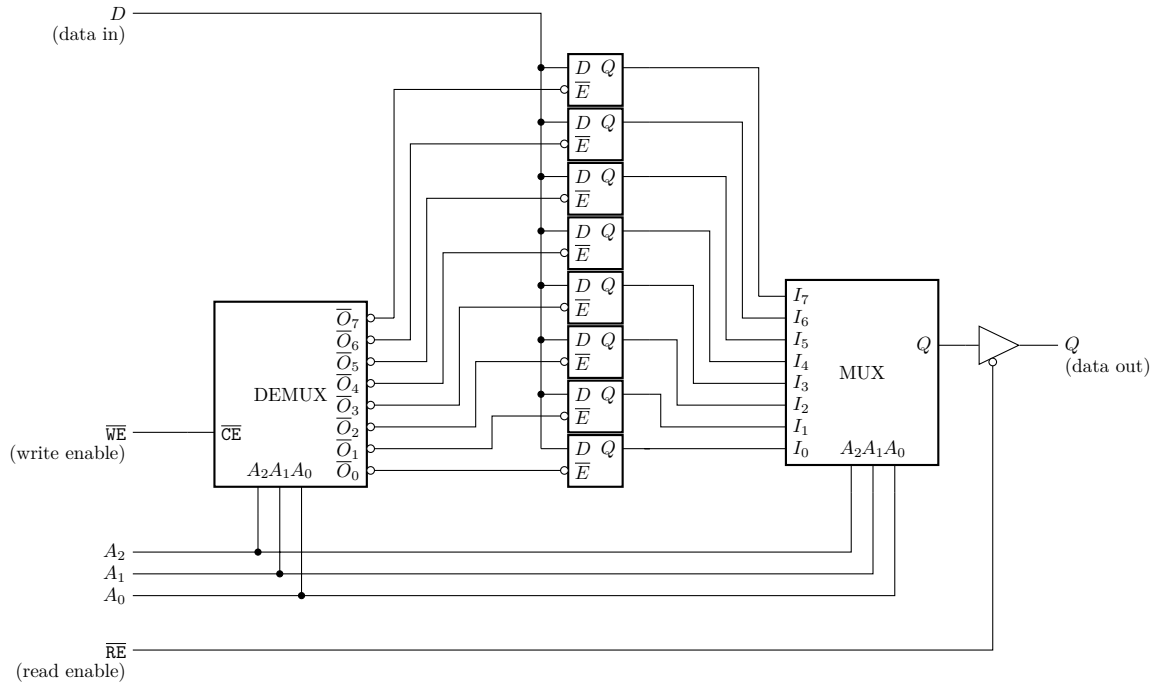
Now, modern computers can have GB of memory on the low end, easily exceeding 10^{10} flip flops. Obviously it would be awkward to have 10^{10} connections to a processor or other device from the memory, so how do we handle this? The answer is to go back to multiplexers and demultiplexers.

13.6.1 Example: 8×1-bit RAM

Below is an example arrangement for a memory circuit with "8×1-bit RAM." This means

- There are 8 "slots" for 1 bit each of memory data (i.e., 1 flip-flop per slot).
- "**RAM**" means **random access memory**, meaning we can easily write and read data to and from any location in memory in any order (as opposed to sequential memory, as on a magnetic tape, or shingled magnetic storage on some modern hard disks).
- Flip-flop based memories like this are called **static RAM (SRAM)**.

The circuit is shown below, with external connections (larger versions of this circuit would come packaged in a single integrated circuit with similar external connections).



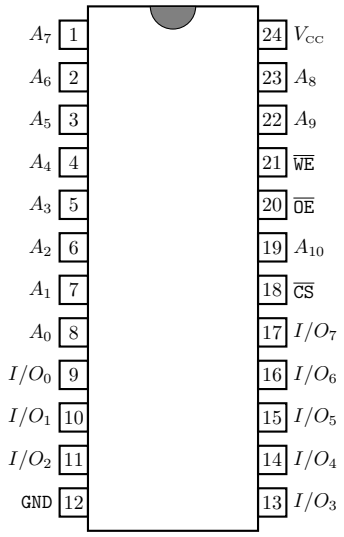
There are a few elements worth noting.

- 8 flip-flops here do the actual storage.
- The address lines A_0–A_2 select which of the 8 flip-flops (slots) is active for either reading or writing.
- For larger memories, the address lines would select a *register* of flip-flops so more data can be transferred in parallel. For example, if the flip-flops were each 4-bit registers, we would have 8×4-bit RAM.
- The data input D is wired to all of the flip-flop inputs; the DEMUX only enables one flip-flop to accept data according to the address.
- The DEMUX must also be enabled by an enable input here $\overline{\text{WE}}$, or write-enable-low. If this input is high, then *no* flip-flop is enabled for data input.

- On the readout side, all flip-flop outputs are fed into a single output via a single MUX, which is addressed by the same address lines as the DEMUX (which is not necessary, but saves address lines if we only want to read *or* write, but not both simultaneously).
- The output Q is buffered by a three-state buffer gate, which is enabled by the $\overline{\texttt{RE}}$ (read-enable-low) input. In this way, several memory chips can share the same output line(s), with only the selected chip attempting to assert a logical value on the shared data output line.

13.6.2 Example: 6116 SRAM

In a example that is more typical than the toy example above, the data input and output lines are the *same*. (Another reason to have a three-state output, so it does not conflict with incoming data.) A old classic, the 6116 SRAM chip, is still available;[3] this is a 2-kB memory (2 kB= 2048×8 bits), and the connections are shown below.



6116 2k×8 SRAM

Most elements are similar to the toy model above.

- The A_0–A_{10} lines form the **address bus**: 11 bits are necessary to address from 0–2047.
- The I/O_0–I/O_7 lines form the **data bus**, which can serve as inputs or outputs (i.e., reading and writing) for the stored data, 1 byte at a time.
- The **chip-select-low** input $\overline{\texttt{CS}}$ (i.e., chip enable) enables the action of the chip when `LOW`. Again, this is useful when several chips share the data bus, so only the enabled chip can write to the bus.
- The **write-enable-low** input $\overline{\texttt{WE}}$ enables the latching of the input data (the I/O lines act as inputs.
- The **output-enable-low** input $\overline{\texttt{OE}}$ enables the data lines to act as outputs.
- If $\overline{\texttt{CS}}$ = `HIGH` or $\overline{\texttt{WE}}$ = `LOW` or $\overline{\texttt{OE}}$ = `HIGH`, then the output buffers are in the high-Z state, again so the data busses of many chips can be connected together. (The address busses are also connected, but these need only act as inputs here.)

[3] For data sheet, see `http://www.idt.com/products/memory-logic/synchronous-and-asynchronous-sram-memory-devices/asynchronous-sram-async-sram/6116-50v-2k-x-8-asynchronous-static-ram`; availability for small quantities starts around $6 each as of May 2015.

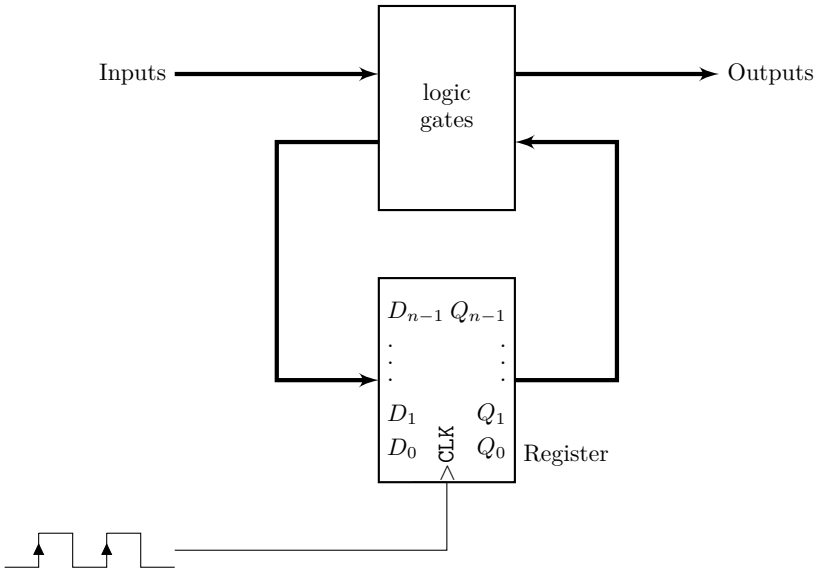
13.6.3 Other Memory Types

The SRAM above is the basic type of digital memory, but there are many other types. We will briefly review them here.

- **DRAM (dynamic RAM)** uses a small capacitor to store a bit of information as a charge state. The disadvantage is that because the capacitor leaks, it must continually be "refreshed" (on ms time scales), which greatly complicates the overhead circuitry. The advantage is that DRAM is cheap and highly scalable; DRAM is standard for large memory modules in modern computers.
- **SRAM (static RAM)** we have already talked about. Why do we want it? It is complicated to fabricate relative to DRAM, and is hard to scale to very large memory. However, it can be relatively power efficient (no refreshing is necessary), and due to the lack of refreshing overhead, it can be much easier to use in small projects. Note that while no refresh is needed, SRAM is volatile (the stored information vanishes when the circuit is not powered).
- **ROM (read-only memory)** is not intended to be written, just read. The data are written during manufacturing.
- **PROM (programmable ROM)** is ROM that can be written (once!) using special programming hardware, which burns fused connections inside the chip by applying relatively high currents.
- **EPROM (erasable PROM)** is PROM that is programmed electronically, and the programming can be erased (usually by exposing the IC to ultraviolet light, through a transparent window in the IC).
- **EEPROM (electronically erasable PROM)** is EPROM that can be erased electronically by the programmer (by applying high electric fields/voltages).

13.7 State Machines with Memory

Before, in Section 13.5, we covered the basic scheme of sequential logic, reproduced below.



Again, the basic idea is to use a register to hold logic values, the outputs of which are transformed and fed back to the register inputs via a logic-array block. Here we will discuss implementing the logic-array block in a very general way by replacing it with memory, either RAM or ROM. In the ROM case, the state machine is suited for fixed operation (e.g., as a counter), whereas with RAM we have the possibility that the state machine can adapt to input and even reprogram itself.

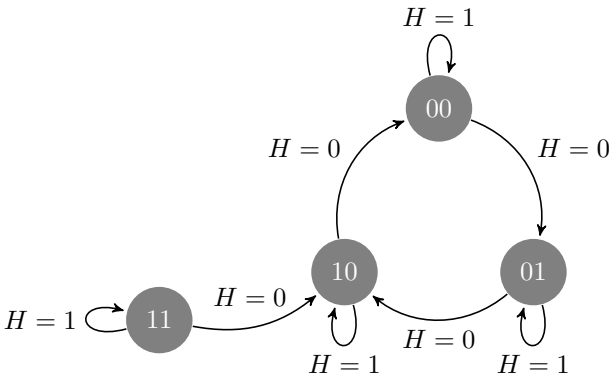
The general idea for implementing state machines with memory is:

- We will connect the register outputs (Q's), representing the *present* state of the machine, to the memory address lines (inputs).
- We will connect the register inputs (D's), representing the *future* state of the machine, to the memory data lines (outputs).
- Any external inputs (needed for the state machine to react to anything external), correspond to extra memory address lines.
- Any external outputs correspond to extra memory data lines.

  It's easiest to see how this works in an example.

13.7.1 Example: Divide-by-3-With-Hold Counter

As an example of a memory-driven state machine, consider the divide-by-3 counter with a hold input H, whose state diagram we considered before in Section 13.5.2.

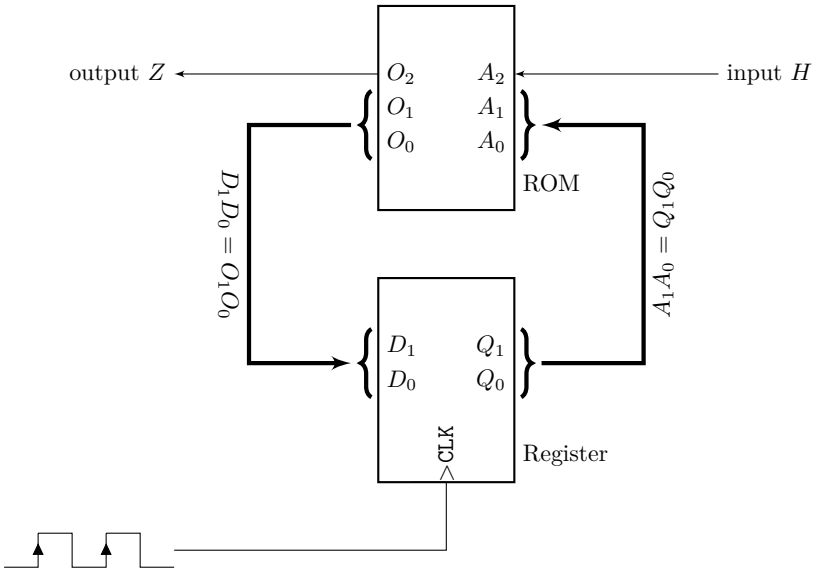


In addition, as an example output bit, we will define a "zero" output bit Z to be `1` if the counter's output bit is `00`, and is `0` otherwise.

The truth table is as follows.

(control)	(present state)		(future state)		(output)
H	Q_1	Q_0	D_1	D_0	Z
0	0	0	0	1	1
0	0	1	1	0	0
0	1	0	0	0	0
0	1	1	0	0	0
1	0	0	0	0	1
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

In this system, there will be 3 address bits (HQ_1Q_2), for 8 memory slots, and 3 data bits (D_1D_0Z), so the size of the memory is $8 \times 3 = 24$ bits. The circuit to implement this is shown below.



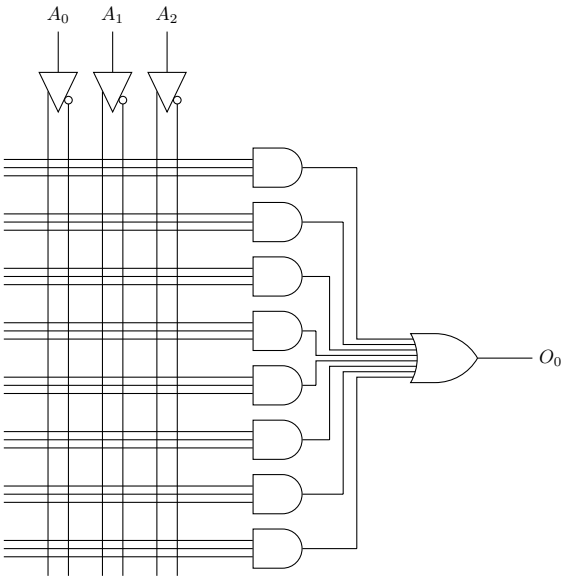
All the logic in the truth table, of course, must be programmed into the ROM. Thus, for example, in address $A_2A_1A_0 = \texttt{000}$, corresponding to $HQ_1Q_0 = \texttt{000}$ in the truth table, we simply program in the value $D_1D_0Z = \texttt{011}$, and so on for the rest of the 8 total memory locations.

13.7.2 General Considerations: Towards a Microprocessor

A few general remarks are in order. First, if there are no input bits, then basically we have some kind of counter (i.e., something that cycles through finitely many states, possibly with some outputs that are a boolean function of the state bits). If there is a single input bit, then it chooses between 2 possible actions (like the hold/count counter). If there are N input bits, then there are 2^N possible operations. This grows quickly with the number of bits: for 8 inputs there are already 256 operations. We can also store sequences of input bits, e.g., in RAM, for effectively many more different possible operations for the same number of input bits (i.e., there could be a 1-bit serial input to a state machine that controls many different possible actions by using different sequences of input bits). These stored bit sequences in RAM correspond to a "program," with 256 "instructions" in this 8-bit example, for a simple realization of a microprocessor. The input/output lines are connected to the data/address busses, which are also connected to input/output devices or interfaces. Real microprocessors often have more specific functionality, sophisticated instruction sets (with instructions that can take multiple clock cycles to complete), and have registers organized in more sophisticated ways than we have indicated here, but we still have the essence of a microprocessor.

13.7.3 Programmable ROM as Logic

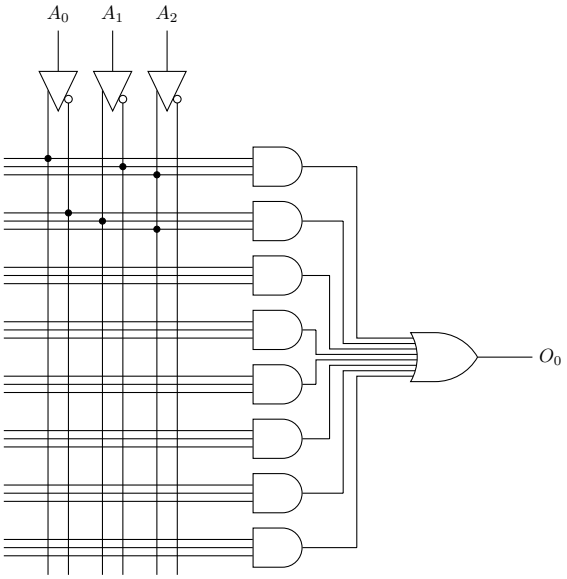
We have talked about replacing logic blocks with ROM, but we can think of ROM itself as a general array of gates, particularly the various forms of PROM. For example, consider the 8×1-bit memory shown below.



The main grid of lines is intended as a configurable grid, where we can make whatever connections we like. Then there are two ways to think of this. First, because the inputs go first into `AND` gates and then into an `OR` gate, we can realize Boolean-logic expressions if they are **sum-of-product** expressions. For example, we can realize

$$O_0 = A_0\overline{A_1}A_2 + \overline{A_0}A_1A_2 \tag{13.2}$$

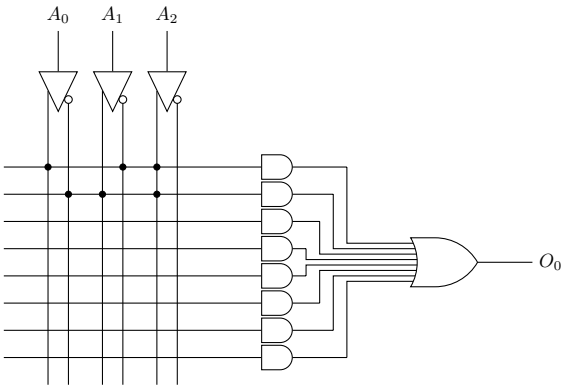
by making the connections shown below.



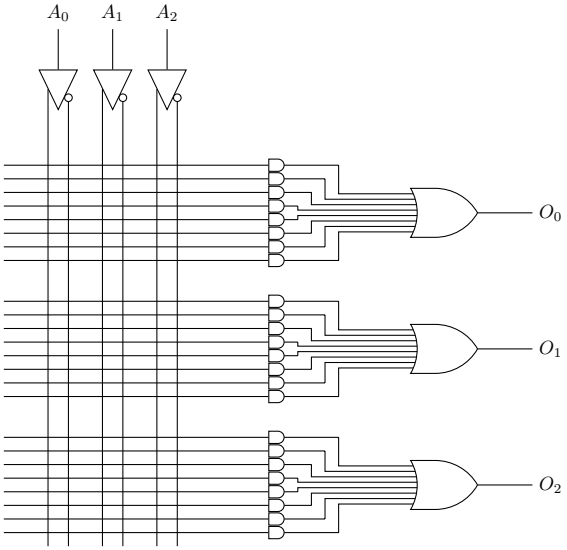
The default of the unconnected inputs is `0`.

The other way to think of this particular example, is that this is a memory that stores the value `1` in the addresses $A_0A_1A_2 =$ `101` and $A_0A_1A_2 =$ `011`, and `0` in all the others. Thus we can store *any* value in any location by making proper connections for all locations that store a `1`. This is why we need 8 `AND` gates in this example—one to "recognize" each possible set of inputs, if needed. (Of course, they are not all used unless all the stored bits are `1`.) In ("write-once") PROM, these connections are all made at the factory, and the undesired connections are "burned" away by flowing high current through the programable fuses at each connection points.

Diagrams like this get to be complicated for larger memories, so they are often abbreviated by "collapsing" all the input lines for a given `AND` gate, as illustrated below. This circuit realizes the same example expression as the previous one.



In this way, we can draw out more complicated memories, like this 3×3 memory.



13.7.4 Programmable Logic Devices

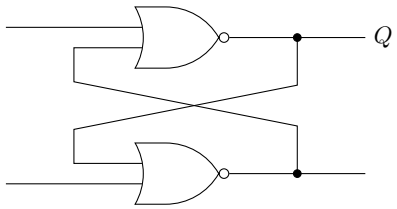
Programmable logic devices (PLDs) are chips that contain a register *and* PROM-type logic arrays like the ones we have shown above. Generally speaking, they do not have sufficient gates to implement *arbitrary* logic combinations (i.e., they generally have fewer than necessary `AND` gates), but they have enough to program a wide range of logic possibilities. Example of simple, but currently available, PLDs are the 22V10 **SPLD**, or **simple PLD** (e.g., the ATF22V10C from Atmel), with 12 dedicated input pins, 10 pins configurable as inputs or outputs, 10 D-type flip-flops, and a gate array (10 `OR` gates, with 10-16 `AND` gates feeding each `OR` gate. The flip-flop outputs can be connected to their corresponding output pins, or the flip-flops can be bypassed altogether for non-registered outputs. A more powerful example is the ATF750C **CPLD**, or **complex PLD** from Atmel, which has the same pin configuration, but provides more gates, and 10 extra "internal" flip-flops that can be used as internal register variables that are not connected directly to outputs.

13.8 Circuit Practice

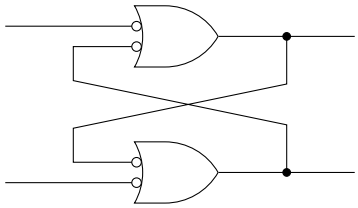
13.8.1 Basic Flip-Flops

For circuit practice, go through these three flip-flops.

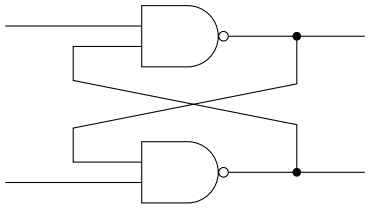
(a) First, label the inputs and remaining output. This is the first one we did, but do this without peeking! Think through the whole truth table.



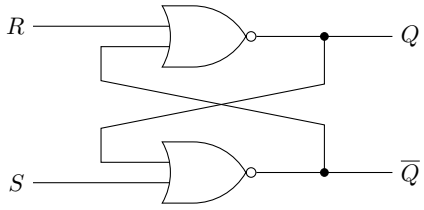
(b) Work out the equivalence of the (a) circuit to this one (i.e., label the inputs and outputs). Try **not** to use a truth table to do this, use a logic theorem to connect these.



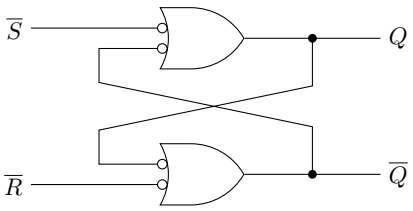
(c) Work out the equivalence of the (b) circuit to this one (i.e., label the inputs and outputs). Try **not** to use a truth table to do this, use a logic theorem to connect these.



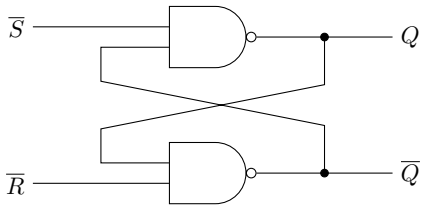
Solution.
(a) The labeled version is:



(b) This is basically the same circuit, but with all inputs and outputs of gates negated. Thus, the flip-flop inputs and outputs are similarly negated. The labeled version is:



Note that since we kept Q in the same spot, we had to swap the inputs as well.
(c) Using $\overline{A} + \overline{B} = \overline{AB}$, we simply change the negated-input `OR` gates to `AND` gates. The labeled version is:



13.8.2 Pulse-Area Stabilizer

We have seen the circuit on the next page before in Section 7.11.5. To review is designed to work as follows. To take photographs with a laser pulse, it it desirable to have the same exposure from each pulse. But the intensity of the laser drifts. Rather than try to stabilize the intensity of the laser, we can compensate for the drift by changing the *duration* of each laser pulse to compensate. By making the **pulse area** or integrated energy of each pulse the same, the photographs have exactly the same exposure, independent of the laser intensity.

Try to trace through the following features in the *digital* part of the circuit.

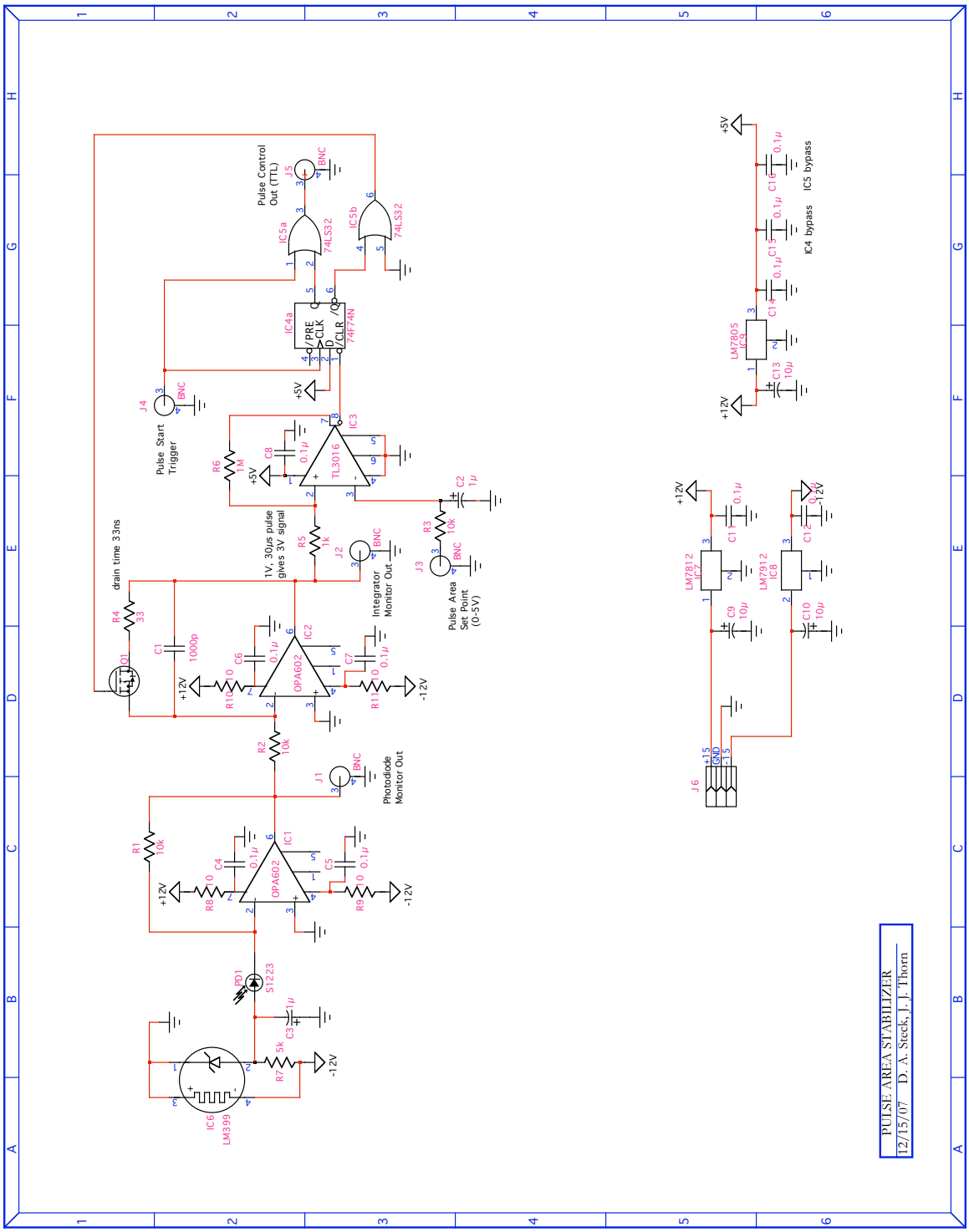
1. At the beginning of the pulse, the "pulse start trigger" input drives the `CLK` input of the D-type flip-flop (IC4a). Since the D input is tied `HIGH`, the Q output goes `HIGH` on the rising edge of this pulse, defining the beginning of the laser pulse.

2. The Q output is `OR`'d with the "pulse start trigger," mainly to allow this input to override the flip-flop state, in case we want the laser to be on continuously for diagnostic purposes.

3. The pulse is finished when the integrated pulse area triggers the comparator IC3, when the inverting output goes low, triggering the $\overline{\texttt{CLK}}$ input of the flip-flop (which triggers on the falling pulse-edge).

4. The $\overline{Q}$ output drives the MOSFET to reset the integrator after the pulse is finished, until the next pulse starts.

5. The propagation delay of IC5a is matched by the propagation delay of the `OR` gate on the $\overline{Q}$ output (IC5b). This is not critical, but we have extra `OR` gates around anyway, and it ensures an accurate $t = 0$ for the integration.


PULSE AREA STABILIZER
12/15/07 D. A. Steck, J. J. Thorn
Pulse Control Out (TTL)
Pulse Start Trigger
drain time 33ns
1V, 30μs pulse gives 3V signal
Integrator Monitor Out
Pulse Area Set Point (0-5V)
Photodiode Monitor Out
IC4 bypass
IC5 bypass
OPA602
TL3016
74LS32
74F74N
LM7805
LM7812
LM7912
LM399
S1223
+12V
-12V
+5V
+15
GND

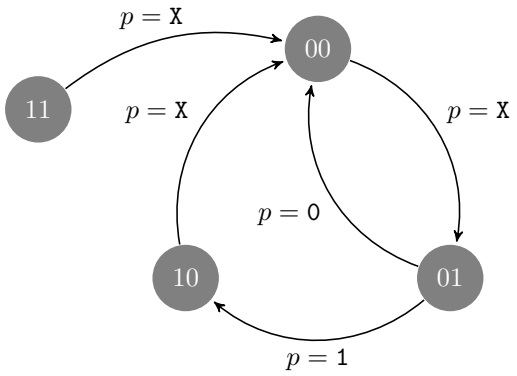
13.8.3 Memory: RAM vs. ROM

Explain the following statement, about connecting memory to a CPU:
With RAM you can scramble the address lines in any order; the same is true of the address lines.
With ROM, you can't!

13.8.4 Circuit Practice: Divide-by-2-or-3 Counter

As practice with state diagrams, draw the state diagram for a divide-by-2-or-3 counter. That is, the counter counts differently based on an input bit p, and the counter counts `00` $\longrightarrow$ `01` $\longrightarrow$ repeat if $p =$ `0`, and `00` $\longrightarrow$ `01` $\longrightarrow$ `10` $\longrightarrow$ repeat if $p =$ `1`. Make sure to handle the excluded state.

Solution. The diagram is sketched below.

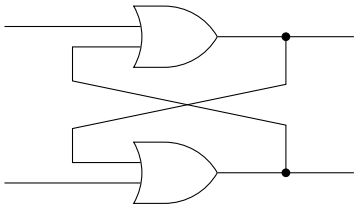


13.9 Exercises

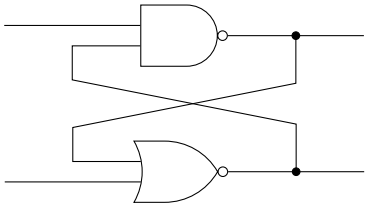
Problem 13.1

(a) What is the essential property of a flip-flop?

(b) Does the circuit below behave as a flip-flop? If so, label the inputs and outputs in flip-flop notation. If not, explain why not.



(c) Does the circuit below behave as a flip-flop? If so, label the inputs and outputs in flip-flop notation. If not, explain why not.

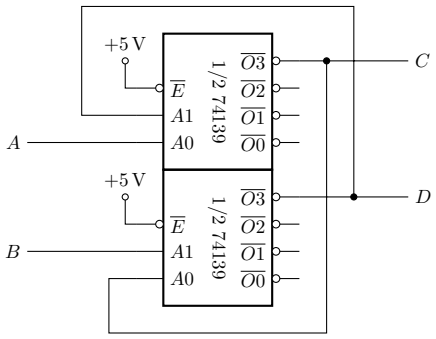


Problem 13.2

Show how to make an RS flip-flop (with both normal and inverted outputs) using an `AND` gate, an `OR` gate, and an `INV` gate. (First try hooking up the `AND` and `OR` as in the usual flip-flop configuration, and then it should be more obvious where the `INV` should go.) Analyze your circuit to show that this behaves as a flip-flop. What is the "bad" input state?

Problem 13.3

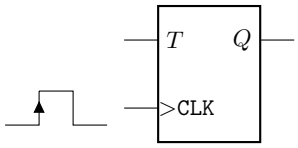
Recall that the 74139 is a 2-bit, 4-output decoder/DEMUX, 2 per package, with 1 inverting enable per decoder. Does the circuit below behave as an SR-type flip-flop (with respect to the labeled inputs/outputs A, B, C, and D)? If so, label the inputs and outputs in flip-flop notation (S, R, Q, $\overline{Q}$). If not, explain how to change the circuit wiring to make it operate as a flip-flop, and label the inputs and outputs in flip-flop notation.



Problem 13.4

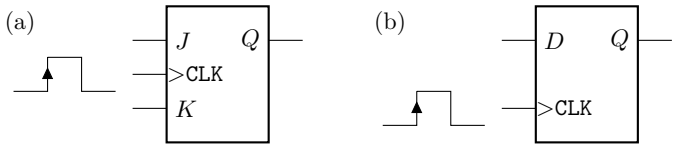
A **T flip-flop** has a single ("T") input, which causes the output Q to toggle if the T = `HIGH`, and to hold if T = `LOW`. The truth table and schematic diagram are shown below (Q_n is the output state Q after the nth clock pulse).

T	Q_n	Q_{n+1}
0	0	0
0	1	1
1	0	1
1	1	0



(a) Show how to connect the JK flip-flop below as a T flip-flop.

(b) Show how to connect the D flip-flop below as a T flip-flop. You will need to use one additional gate.



Problem 13.5

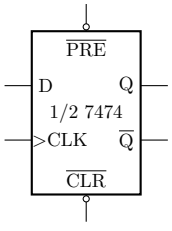
(a) Show how to connect 3 JK flip-flops to make a 3-bit ripple counter, and then show how to add a single 2-input `NAND` gate to change this to a **modulo-5** counter—that is, the counter resets to zero when the output reaches 5, and continues counting.

(b) Show how to design a (ripple) counter that counts from 0 to 5 and then stops. Your circuit should include a $\overline{\texttt{RESET}}$ input that resets the counter back to zero (and then continue counting) after a negative pulse.

(c) Show how to design a circuit that passes only 5 (positive) pulses and blocks subsequent other pulses, after a negative reset pulse that "arms" the circuit.

Problem 13.6

(a) Show how to connect 3 flip-flops to make an asynchronous (ripple), 3-bit **down** counter. To be specific, use 7474 flip-flops (dual D-type, positive-edge-triggered, with complementary outputs and jam preset and clear), whose connections are shown below.

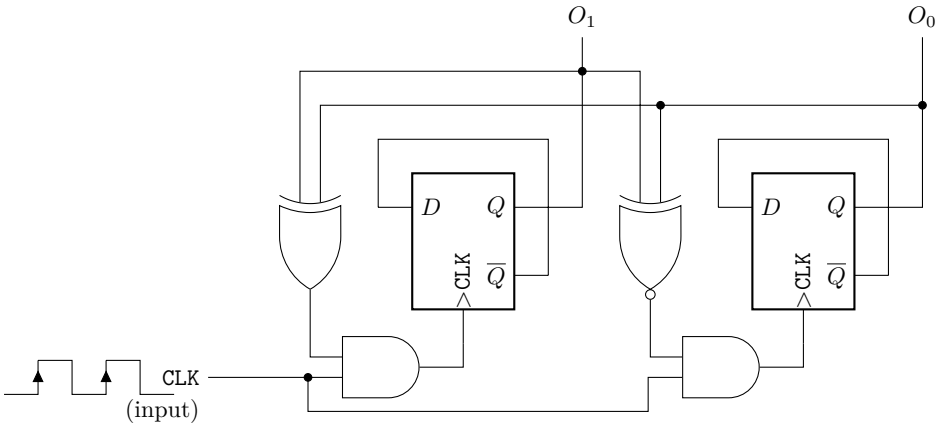


Recall that for the jam inputs, "preset" is the same thing as "set," and "clear" is the same thing as "reset."

(b) Show how to realize an **asynchronous, divide-by-5 down counter**, made from the same D-type flip flops. That is, your counter should count 4, 3, 2, 1, 0, 4, 3, 2, 1, 0, ...

Problem 13.7

Consider the asynchronous counter circuit shown below.



(a) Why is this circuit "asynchronous" and not "synchronous"?

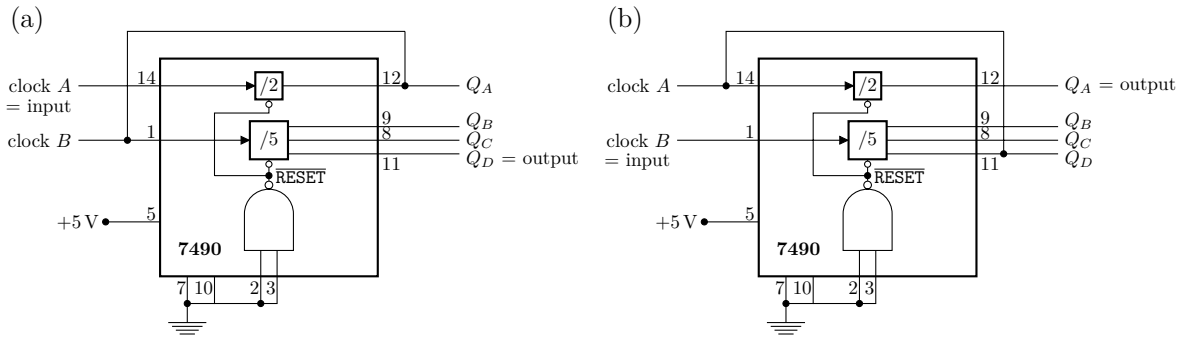
(b) Assuming the counter starts in the state $O_1O_0 = \texttt{00}$, give the truth table for the counter output on subsequent clock cycles.

(c) Can this counter get "stuck" in any particular state? Why or why not?

(d) What kind of counter is this?

Problem 13.8

Recall that the 7490 is a decade counter, with clock inputs that trigger on falling edges. Below are two ways to connect the 7490 as a divide-by-10 counter. However, they are not equivalent. What, specifically, is the difference in the output waveforms? (Be **quantitative**.) Also, remember Q_B is the least significant bit of the divide-by-5 subcounter.



Problem 13.9

Design a synchronous 2-bit UP/DOWN counter: It has a clock input, and a control input ($U/\overline{D}$); the outputs are the two flip-flop outputs Q_1 and Q_2. If $U/\overline{D}$ is `HIGH`, it goes through a normal binary counting sequence; if `LOW`, it counts backward—$Q_2Q_1 = \texttt{00}, \texttt{11}, \texttt{10}, \texttt{01}, \texttt{00}, \ldots$[4]

Problem 13.10

Design a **synchronous**, 3-bit Fibonacci counter (i.e., count through the Fibonacci numbers 0, 1, 2, 3, 5, and then repeat). Use three flip-flops (of the 7474 type, as shown above) and whatever gates you

[4]Paul Horowitz and Winfield Hill, *The Art of Electronics*, 2nd ed. (Cambridge, 1989), Exercise 8.25 (ISBN: 0521370957).

like. Be sure to show a state diagram, *including* all excluded states. Can your counter get "stuck" in any excluded state?

Problem 13.11

Design a synchronous divide-by-3 circuit using two JK flip-flops. It can be done (in 16 different ways) without any gates or inverters. One hint: When you construct the table of required J_1, K_1 and J_2, K_2 inputs, keep in mind that there are two possibilities for J, K at each point. For instance, if a flip-flop output is to go from 0 to 1, $J, K = 1, \texttt{X}$ ($\texttt{X}$ = doesn't matter). Finally, check to see if the circuit will get stuck in the excluded state (of the 16 distinct solutions to this problem, 4 will get stuck and 12 won't).[5]

Problem 13.12

(a) (15 points) Design a **synchronous** circuit (state machine) using flip-flops and logic gates that makes a 4-bit "Knight-Rider" pattern. That is, the output counts:

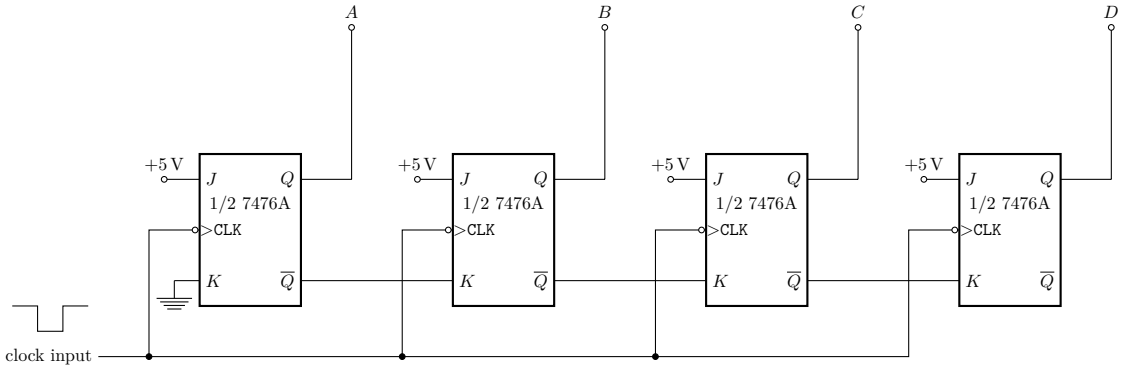
```
0001
0010
0100
1000
0100
0010
(repeat).
```

Don't draw the circuit, just come up with the required logic expressions. *Hint:* there are 4 bits shown here, but you will need an extra bit to keep track of the direction.

(b) (10 points) Of course, if the Knight Rider car is driving down the street with the lights stuck in an excluded state, it would be less than awesome. Either show that your circuit doesn't get stuck in excluded states, or show how to modify your circuit to make sure that your circuit settles into the above pattern for any starting state. Use whatever logic you like (again, no circuit, just the required logic).

Problem 13.13

Analyze the circuit below by drawing a timing diagram for the outputs. Assume that the initial state is $ABCD = \texttt{0111}$, and analyze the circuit output for 5 clock pulses.



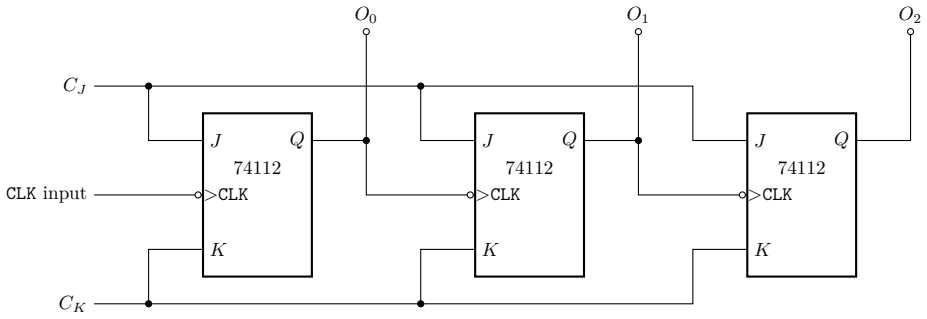
Problem 13.14

Design a **synchronous**, divide-by-16 counter. Just work out Boolean expressions for the required logic using the design procedure for the state machine; no need to draw a circuit schematic. (Write your output bits as $Q_3Q_2Q_1Q_0$ from MSB to LSB.)

[5] Paul Horowitz and Winfield Hill, *op. cit.*, Exercise 8.24.

Problem 13.15

Consider the ripple-counter circuit shown below, with control inputs C_J and C_K.



What control-input combination C_JC_K is necessary for the circuit to count, and what kind of counter is it? Describe the operation of this circuit for all other control-input combinations C_JC_K.

Problem 13.16

(a) Design a **synchronous** 4-bit, binary counter, using JK flip-flops (use the same 74112's as shown in Problem 15), by considering the following: Think about how binary counting works, and in particular, what is the condition for a particular bit to toggle its state during the counting sequence? How do you switch between toggling or holding in a JK flip-flop?

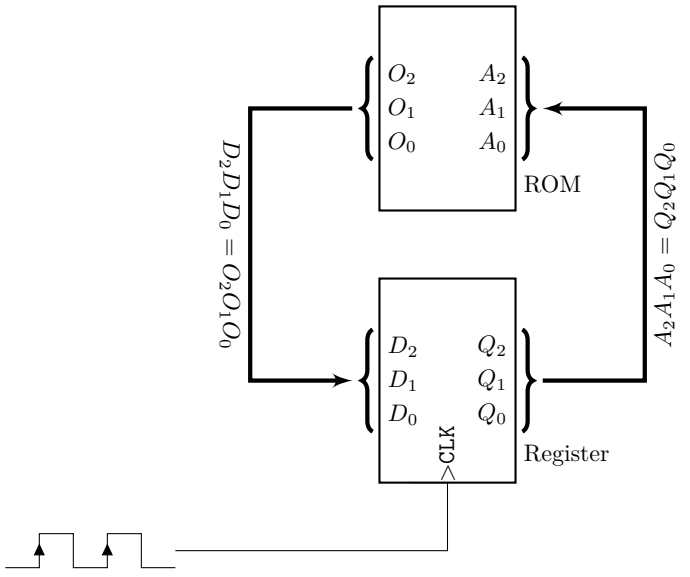
(b) Design a **synchronous** 4-bit, binary counter, as in part (a), but this time using D flip-flops. Use the same hint from part (a). What kind of gate acts as a conditional inverter?

Problem 13.17

Design a circuit that combines 4 of the 6116 SRAM IC's (2k×8-bit) to make a single, 4k×16-bit memory. The resulting circuit should have the same behavior as one of the original 6116's, just with more address/data bits.

Problem 13.18

A 3-bit register and an 8×3 ROM are connected as shown below.



The ROM is programmed as follows:

A_2	A_1	A_0	O_2	O_1	O_0
0	0	0	1	0	0
0	0	1	0	1	0
0	1	0	0	0	0
0	1	1	0	1	1
1	0	0	1	1	0
1	0	1	0	0	0
1	1	0	1	1	1
1	1	1	0	0	1

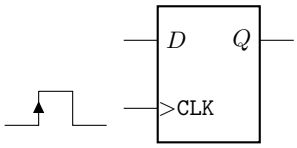
Assume that the initial state of the register is $Q_2Q_1Q_0 = 010$.

(a) Give the state of the register outputs $Q_2Q_1Q_0$ after *each* of the next three clock pulses.

(b) Even after arbitrarily many clock pulses, (at least) one possible value of $Q_2Q_1Q_0$ will never occur. Which?

Problem 13.19

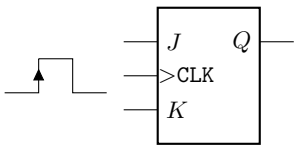
Show how to connect a T flip-flop as a D flip-flop (shown below). (The operation of the T flip-flop is defined in Problem 4.) Use whatever extra logic gates you need (simple two-input gates only).



Hint: if you don't see how to get started, treat the problem as a state machine with current state Q_n and external input D.

Problem 13.20

Show how to connect a T flip-flop as a JK flip-flop (shown below). Recall that the operation of the T flip-flop is defined in Problem 4. You can use whatever extra logic gates you need (simple two-input gates only).



For this problem, follow the outline below.

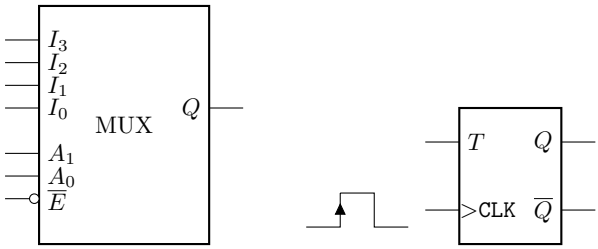
(a) Begin the design as a state machine by making a truth table, with columns: J, K, and Q_n, Q_{n+1}, and T. (The J, K, and Q_n variables are the inputs and current state, Q_{n+1} the future state, and T the flip-flop input where you need to supply the appropriate logic function as input.)

(b) Make a Karnaugh map for the T variable in terms of the input variables and current state, and use it to write down a simple logic expression for T.

(c) Complete the design by making a schematic diagram that realizes the JK flip-flop in terms of the T.

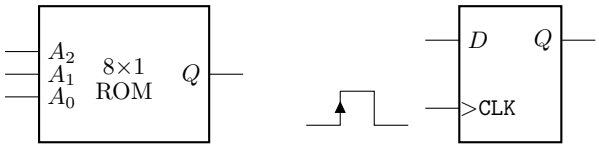
Problem 13.21

Show how to connect a T flip-flop as a JK flip-flop, but using only an extra 4-input MUX, and *no extra logic gates*. The T flip-flop and MUX you should use are shown below. (The operation of the T flip-flop is defined in Problem 4.)



Problem 13.22

Show how to connect an 8×1 ROM (shown below) as a JK flip-flop. You will need an extra D flip-flop (also shown below), but *no other logic gates*. Also specify how the ROM should be programmed to make your circuit work properly.



Chapter 14

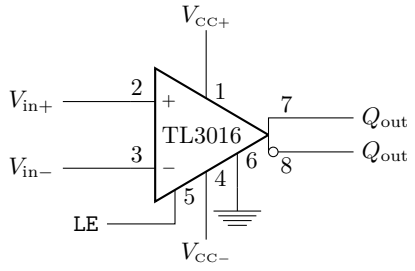
Comparators

14.1 Overview and Review

We talked before about comparators in the context of analog electronics in Section 7.9. There, comparators were variations on the basic op-amp, which compares two analog input voltages, and outputs something like a `1` or `0`, depending on the comparison. Here, we will review and extend our discussion of comparators, because in the context of digital electronics, comparators are the fundamental way to interface analog to digital circuits. For example, they form the fundamental building block of the **analog-to-digital converter (ADC)**. They also represent a basic method for generating logic pulses (generally, the pulse timing is based on an analog signal, such as an RC decay). Also, comparators allow you to do "level translation" between different logic types (e.g., interfacing high-voltage logic to TTL).

14.1.1 Example: TL3016

As an example of a comparator that is really optimized for driving logic circuits, consider the TL3016 (or LT3016) fast comparator ("fast" here means a 7.6-ns propagation delay). This comparator and its pin connections are shown schematically below.



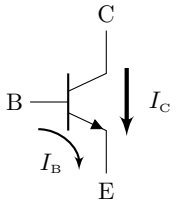
This comparator has both regular and inverted outputs, and the outputs are TTL-compatible. To summarize the regular operation:

- if $V_{in+} > V_{in-}$: $Q =$ `HIGH`, and $\overline{Q} =$ `LOW`
- if $V_{in+} < V_{in-}$: $Q =$ `LOW`, and $\overline{Q} =$ `HIGH`
- here, "`HIGH`" means nominally around +5 V (actually, about +3.8 V)
- and "`LOW`" means nominally around 0 V (actually, about +0.6 V)
- for proper operation, the inputs should be in the range of the supply voltages; the positive supply V_{CC+} is a +5 V, while the negative supply V_{CC-} is either 0 V or −5 V
- the "`LE`" pin is a "latch enable", which latches the output when held `HIGH`

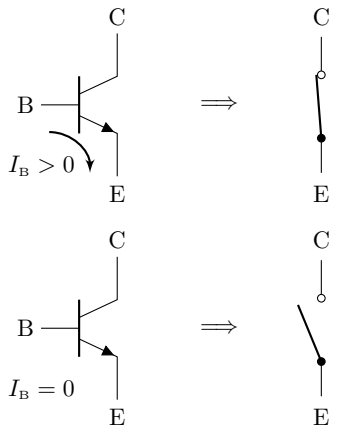
14.2 Open-Collector Output

A common configuration for comparator outputs (and logic-gate outputs, too) is the **open-collector output**. We discussed this before in Section 7.9, but we will review the basic idea here.

First, we should review the basic switch-type operation of the bipolar transistor. The transistor acts as a switch for current, based on another current. The two important currents are I_B from the base to the emitter, and I_C from the collector to the emitter.

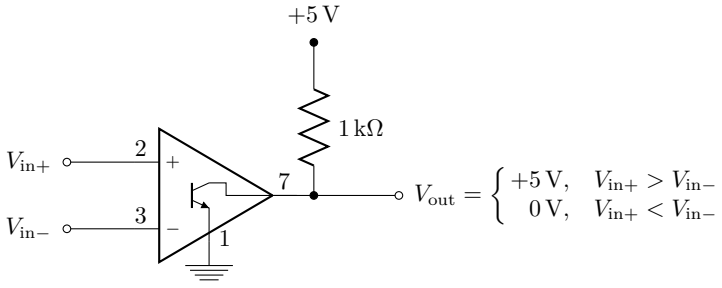


The base current I_B acts as the control current, and I_C is the current to be switched. Simplistically, if there is some current I_B, then I_C can flow, so the C–E path acts as a closed switch.



However, if $I_B = 0$, then the C–E path acts as an open switch. There are some extra voltage drops to consider here, but as in RTL logic (Section 11.3), this simple model is sufficient to understand open-collector outputs.

An inexpensive an popular comparator with open-collector output is the LM311 (also LF311, henceforth just the "311"). Schematically, this comparator is shown below, with a typical "pull-up resistor" to +5 V on the output.

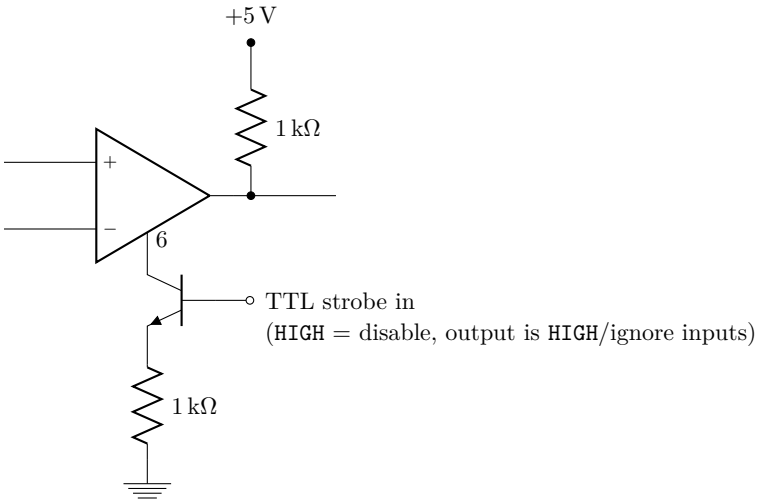


Not shown here are power-supply connections (to supplies of up to ±18 V). The inputs must stay within the supply-voltage range. To summarize the operation here:

- if $V_{in+} > V_{in-}$: the transistor is `OFF`, and $V_{out} = +5\,V$
- if $V_{in+} < V_{in-}$: the transistor is `ON`, and $V_{out} = 0\,V$ (actually, about 0.2 V or higher)

The point of the open-collector output is its flexibility: it's not restricted to particular voltages like the outputs of normal TTL gates. For example, the 311 can drive loads up to 40 V and 50 mA, so it can directly control LEDs, lamps, relays, and so on. By contrast, a TTL-gate outputs would need a buffer transistor.

Another feature of the 311 is a TTL strobe input. A typical connection for this pin is shown below.

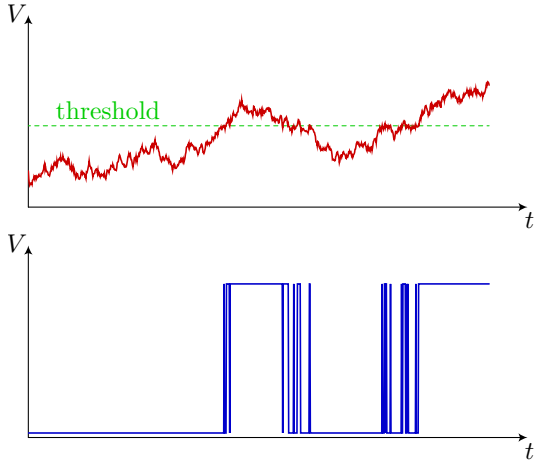


If the strobe input is held `HIGH`, then the output is disabled: the output is `HIGH` (i.e., the output transistor opens), and the comparator ignores the inputs. This is useful for "gated" operation, if the comparator should only trigger during some time interval or some condition determined by other logic.

14.3 Schmitt Trigger

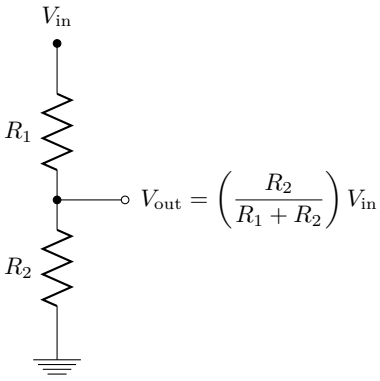
We also discussed the **Schmitt trigger** in the context of analog circuits before in Section 7.9.1. We will review the basic idea again here, since these are so important in digital applications that many logic devices have integrated Schmitt triggers.

The motivation for the Schmitt trigger comes from noisy inputs signals. Typically, these are slowly changing analog signals, but even "digital" signals are fundamentally analog, and are thus similarly susceptible to noise. As illustrated below, a noisy signal that is rising and falling can cause many spurious transitions while crossing the threshold.

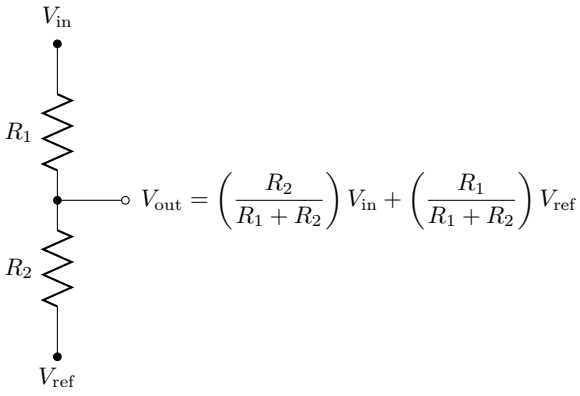


Ideally, with a noise-free signal, the output (shown in the bottom plot) would have one up transition, then later a down transition, and then an up transition again, instead of the many clustered around the three "ideal" trigger times.

To analyze the fix for this, we will briefly review the voltage divider (see Section 1.3.3 and Problem 2). Given a series pair of resistors supplied by V_{in}, the output voltage at the tap point is given in terms of the "fractional resistance" at the tap point.



If there are two voltages at either end of the resistor pair, the output voltage is a linear combination of the two voltages, given by the fractional resistances:

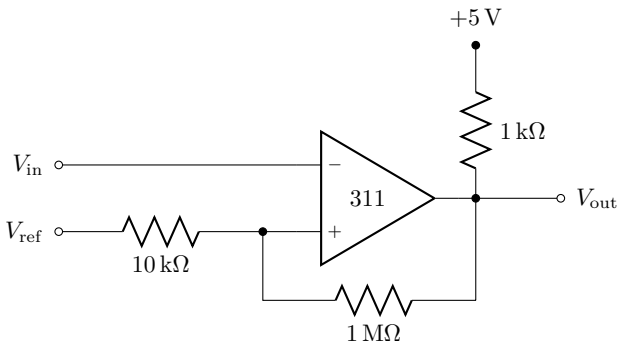


We can deduce the second result from the first by subtracting V_{ref} from all voltages, and applying the first (grounded) result, to give

$$V_{\text{out}} - V_{\text{ref}} = \left(\frac{R_2}{R_1 + R_2}\right)(V_{\text{in}} - V_{\text{ref}}). \tag{14.1}$$

Rearranging gives the result in the figure. *Note:* you should *memorize* both of these voltage-divider formulas, and be able to quickly come up with resistor combinations that divide a voltage by 2, 3, etc.

Now consider the following circuit, which is a comparator with two added resistors. One resistor is in series with the "trigger voltage" V_{ref}, and another, large resistor ties the output to the noninverting input.



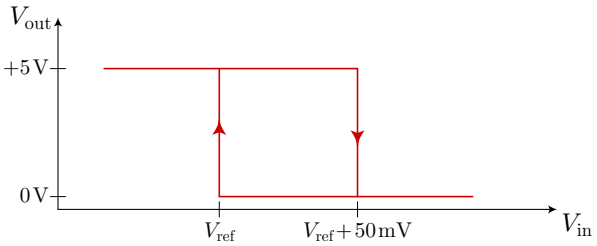
The extra resistors introduce hysteresis into the operation of the comparator as follows.

- if $V_{\text{out}} = 0$, V_+ (the voltage at the noninverting input) is $0.99\,V_{\text{ref}}$, using the first voltage-divider formula.
- if $V_{\text{out}} = +5\,\text{V}$, $V_+ = 0.99\,V_{\text{ref}} + 0.01 \cdot 5\,\text{V}$, using the second voltage-divider formula.

Thus, the reference voltage changes by about 50 mV, depending on the output. Notice that

- if $V_{\text{in}} = \texttt{HIGH}$, the trigger point is *lower*
- if $V_{\text{in}} = \texttt{LOW}$, the trigger point is *higher*

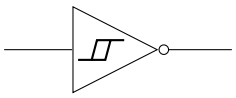
So V_{in} "repels" the trigger point, which gives immunity to noise. The hysteresis of the Schmitt trigger is sketched in the output-response plot below.



Note that the Schmitt trigger is bistable in the 50-mV-wide region near the nominal trigger voltage V_{ref}. Thus, once the Schmitt trigger makes a transition, noise of less than around 50 mV will not cause another spurious transition.

Note that the configuration above is inverting, since V_{in} goes into the inverting input (i.e., large V_{in} means the output is `LOW`). For a noninverting comparator, you can swap the V_{in} and V_{ref} labels, but note that the input is no longer well-isolated from the input, which may be a problem if the input source has high impedance.

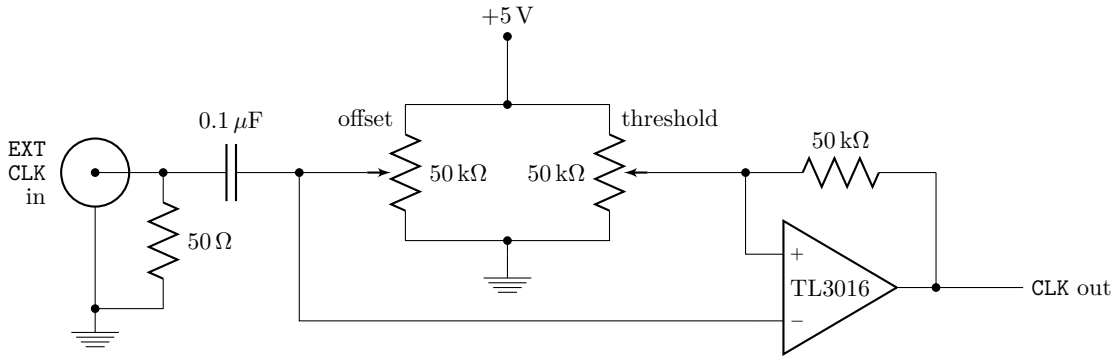
In digital circuits, many gates and logic devices are available with Schmitt-trigger inputs. An example is the 7414, a hex Schmitt-trigger inverter. This is shown schematically below.



Note the Schmitt-trigger symbol on the gate, which suggests the hysteresis curve. Schmitt-trigger-input gates are good for signals without well-defined edges (like a sinusoidal input), or signals from an external source with a long cable run, which could be susceptible to noise pickup.

14.3.1 Example: Analog-to-Digital Clock-Signal Conversion

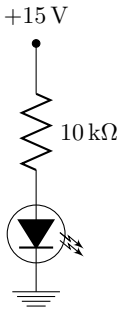
As an example application of a Schmitt trigger, consider the conversion of an analog clock signal to a digital clock. An example of an analog clock source is a rubidium atomic clock, a relatively inexpensive (few $K) instrument that typically provides a 10-MHz sine wave, with a stability/accuracy of better than a part in 10^{10}. To drive digital circuits, however, this should be converted into an appropriate square wave with logic-level inputs. The circuit below accomplishes this, using a Schmitt trigger to avoid spurious extra clock pulses due to noise on the clock signal.



Note the ac-coupled input, a 50-Ω-terminated input (as appropriate for a 50-Ω cable connecting the clock to this circuit), and the Schmitt trigger, with TTL-compatible output. The clock offset and threshold voltages are adjustable to account for different input clock amplitudes, and to adjust the duty cycle of the resulting square wave (typically, these can both be set to 2.5 V.

14.4 Circuit Practice

For practice with comparators, first note that you can light an LED with a power supply and a current-limiting resistor as follows.

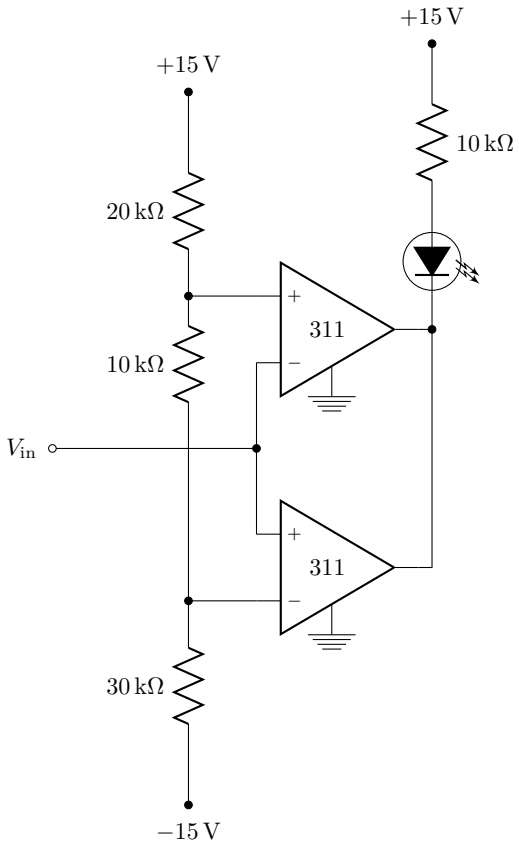


Thus, an open-collector output can control the LED, since the output is either open or shorted to ground (corresponding to an `OFF` or `ON` LED, respectively).

Now design a "TTL out-of-range alarm," given the following components and requirements:

- two 311's
- any resistors you like
- an LED
- ±15-V power supplies (and ground)
- the circuit operates as follows: if $V_{\text{in}} > +5\,\text{V}$ or $V_{\text{in}} < 0\,\text{V}$, the LED "alarm" lights; otherwise, the LED is off

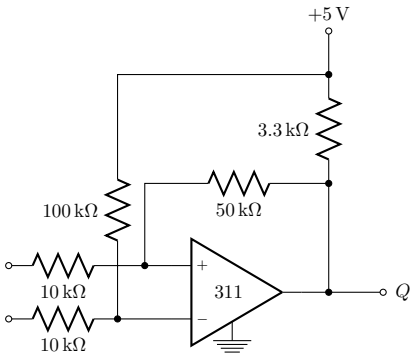
Solution.



14.5 Exercises

Problem 14.1

(a) Show that the comparator circuit below acts as a flip-flop ("**bistable multivibrator**"). Label the set and reset inputs.



(b) Show how to add another (311) comparator to also produce the $\overline{Q}$ output. (Use whatever resistors you like.)

Problem 14.2

Briefly describe how a Schmitt trigger in an input of a logic gate can improve the performance of a circuit. Under what conditions on the input signal do you expect an improvement?

Chapter 15

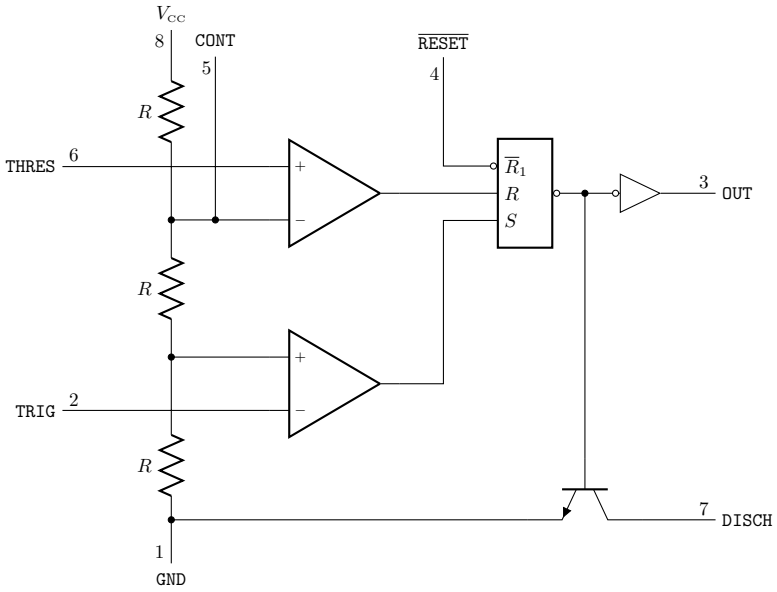
Pulse and Waveform Generation

15.1 The Classic 555 Timer

The 555 timer is an old, classic workhorse for timing and pulse-signal generation, in applications that require moderate accuracy (1%) and relatively slow signals (good performance up to a few hundred kHz, and can be pushed up to $\sim$1 MHz. It's a versatile chip, and can produce square waves, arbitrary-length pulses, and can perform more complicated tasks such as pulse-width modulation—there is a lot of functionality packed into this 8-pin chip.

15.1.1 Equivalent Circuit

The functional equivalent circuit for the 555 chip is shown below, with pin assignments for an 8-pin DIP package.[1]



A few things to notice here:

1. V_{CC} powers all the components in the chip. The output is TTL compatible when $V_{CC} = +5$ V, but V_{CC} can go up to +18 V. The minimum for the standard chip is +4.5 V, but some CMOS variants can make use of lower voltages (e.g., +2 V for the ICL7555, +1 V for the TLC551). The `GND` (ground) input sets the ground reference for the circuit.

[1] For a nice view inside the package and discussion of the "guts" of the 555, see `http://www.righto.com/2016/02/555-timer-teardown-inside-worlds-most.html`.

2. V_{CC} also drives a resistor divider chain, which sets voltage-reference points at $(2/3)V_{CC}$ and $(1/3)V_{CC}$. The `CONT` (control) input taps into the $(2/3)V_{CC}$ point, and can be used as an "override" for the reference voltages.

3. Two comparators compare two input voltages (`THRES` or threshold, and `TRIG` or trigger) to these reference voltages.

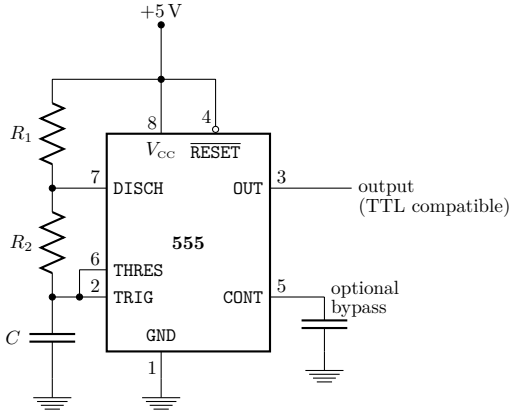
4. The comparator outputs drive the inputs of an SR flip-flop. The flip-flop also has an externally connected, direct-reset input ($\overline{\texttt{RESET}}$).

5. The output of the flip-flop is buffered and set to the `OUT` port. The standard 555 (NE555) has a lot of "oomph," and can handle ± 200 mA of current. Note that "setting" the flip-flop takes the output `HIGH`.

6. The flip-flop output also connects to the base of an open-collector, NPN transistor, whose collector is the `DISCH` (discharge) output. This output is useful, for example, for dumping the charge of a timing capacitor. In this case, "resetting" the flip-flop will turn the transistor on (i.e., capacitor charge gets dumped), setting the flip-flop will turn the transistor off.

15.1.2 Astable Multivibrator

A typical 555 circuit is shown below. This is called an **astable multivibrator**, which just means that the output is a square wave, whose timing is set by the external components.



Let's analyze how this works:

1. The bypass capacitor at pin 5 (`CONT`) helps to stabilize the reference voltages in the resistor chain. One problem with the 555 is that it can cause large power-supply transients while switching, which can feed into the reference chain and cause multiple (unintended) transitions and glitching, and a capacitor here helps to fight these problems. It's a good idea to bypass pin 8 (V_{CC}) with a large capacitor as well.

2. The main timing of the square wave is controlled by charging and discharging the capacitor C. The capacitor charges from V_{CC} via $R_1 + R_2$, while the capacitor discharges only via R_2 to pin 7, when the transistor is on and this pin is connected to ground.

3. We will assume that the capacitor is initially uncharged. Note that as long as `THRES` (pin 6) is below $(2/3)V_{CC}$, the corresponding comparator output is `LOW`, which does not reset the flip-flop. The other comparator sets the flip-flop since the capacitor voltage at pin 2 (`TRIG`) is below $(1/3)V_{CC}$. This sets the output `HIGH` and turns the transistor off, so the capacitor charges via $R_1 + R_2$.

4. When the capacitor voltage reaches $(2/3)V_{CC}$, the flip-flop gets reset by the `THRES` comparator, setting the output `LOW` and turning the transistor on, shorting pin 7 (`DISCH`) to ground.

5. The capacitor discharges through R_2 until the voltage drops to $(1/3)V_{CC}$, when the flip-flop sets again, and the process repeats.

6. For either phase, we can use the exponential relaxation of the RC circuit to figure out the times in each state. For the output `LOW`-output phase, the capacitor voltage relaxes exponentially (with time R_2C) from $(2/3)V_{\rm CC}$ towards 0 (neglecting the small collector-emitter voltage across the discharge transistor). That is, the capacitor voltage is
$$V(t) = \frac{2}{3}V_{\rm CC}\, e^{-t/R_2C}, \tag{15.1}$$
and if we set $V(\tau_{\rm low}) = (1/3)V_{\rm CC}$, where $\tau_{\rm low}$ is the `LOW`-cycle time, we get
$$\tau_{\rm low} = (\log 2)R_2C \approx 0.693\, R_2C. \tag{15.2}$$
Note that "log" here is the natural logarithm.

7. For the output `HIGH`-output phase, the capacitor voltage charges exponentially [with time $(R_1 + R_2)C$] from $(1/3)V_{\rm CC}$ towards $V_{\rm CC}$. Note that, except for the time constant, this is the same as the `LOW` phase, except for an inversion and a shift in voltage. Thus, a similar result applies, with
$$\tau_{\rm high} = (\log 2)(R_1 + R_2)C \approx 0.693\, (R_1 + R_2)C \tag{15.3}$$
as the high-time dwell state. Note that $\tau_{\rm high} > \tau_{\rm low}$ (note that we can't have $R_1 = 0$, otherwise the discharge transistor would attempt to short $V_{\rm CC}$ to ground—not a good idea.

8. The oscillation period is then
$$T = \tau_{\rm low} + \tau_{\rm high} = (\log 2)(R_1 + 2R_2)C \approx 0.693\, (R_1 + 2R_2)C. \tag{15.4}$$
Note that $V_{\rm CC}$ dropped out of this expression, which is convenient, since it means the timing is insensitive to the power-supply voltage. Of course, we assumed that the power supply is *constant*; if $V_{\rm CC}$ varies on the time scale of the period, such that the voltage is differs between successive transitions, then $V_{\rm CC}$ *doesn't* drop out of the period. Hence, again, the importance of a bypass capacitor for the power supply.

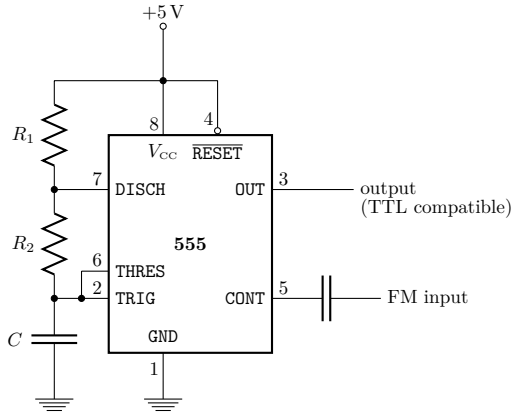
9. Of course, there are some limitations to what is possible with the above timing. Lancaster[2] recommends: $R_1 + R_2 \leq 3.3\,{\rm M}\Omega$, $R_1, R_2 \geq 1\,{\rm k}\Omega$, $C \geq 500\,{\rm pF}$. The capacitor can be large, but the RC times may ultimately be limited by capacitor leakage, but hours-long periods are possible.

10. A nearly symmetric square wave results if $R_1 \ll R_2$, but if the asymmetry is a problem, the oscillator frequency can be doubled, and the output fed through a divide-by-two circuit, resulting in a symmetric wave, independent of the initial asymmetry (why?).

15.1.2.1 Frequency Modulation

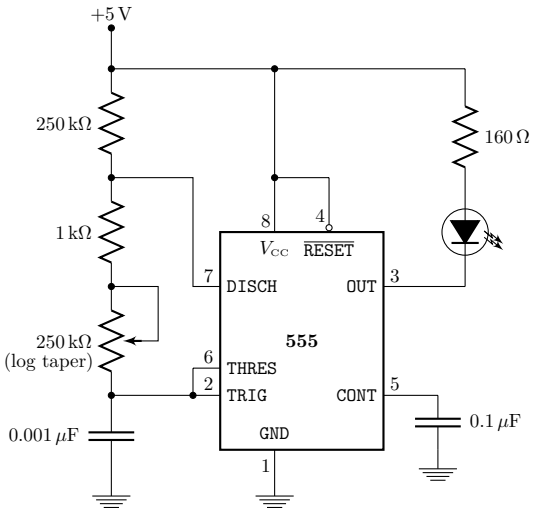
A variation on this circuit is to use the `CONT` input to modulate the $(2/3)V_{\rm CC}$ reference point, which modulates the frequency (raising the reference voltage = longer time to trigger the comparator = lower frequency). A coupling capacitor makes the frequency modulation more convenient, as a zero voltage corresponds to no modulation.

[2] Don Lancaster, *The TTL Cookbook* (SAMS/Prentice Hall, 1974), p. 173.



15.1.2.2 Pulse-Width Modulation: LED Dimmer

The asymmetry of the 555 output is sometimes useful, as in **pulse-width modulation**. You can control the brightness of an LED by controlling the supply current; however, if the LED is driven by a fixed-voltage, digital output, this isn't feasible. The solution is to blink it rapidly on and off, and vary the fraction of time it is on (i.e., vary the **duty cycle**—the fraction of the period where the signal is `HIGH`—of the pulse). The 160 Ω resistor here sets the LED current to 20 mA, assuming a 1.8-V drop across the LED (as appropriate for a standard red LED), and assuming the output goes down to 0 V (almost true). The component values give a minimum frequency of about 1.3 kHz (plenty fast to make the LED appear continuous—it should be over about 50 Hz), and the duty cycle varies from a maximum of about 50% down to a minimum of about 0.4%. (Note that the LED is on when the output is `LOW`.) This is also an *efficient* dimming scheme, which is why it is so common: restricting the current generally requires some wasted power, due to power dissipated in a current-limiting resistance. In this circuit, the efficiency is roughly the same at any value (neglecting power dissipated due to output transitions).

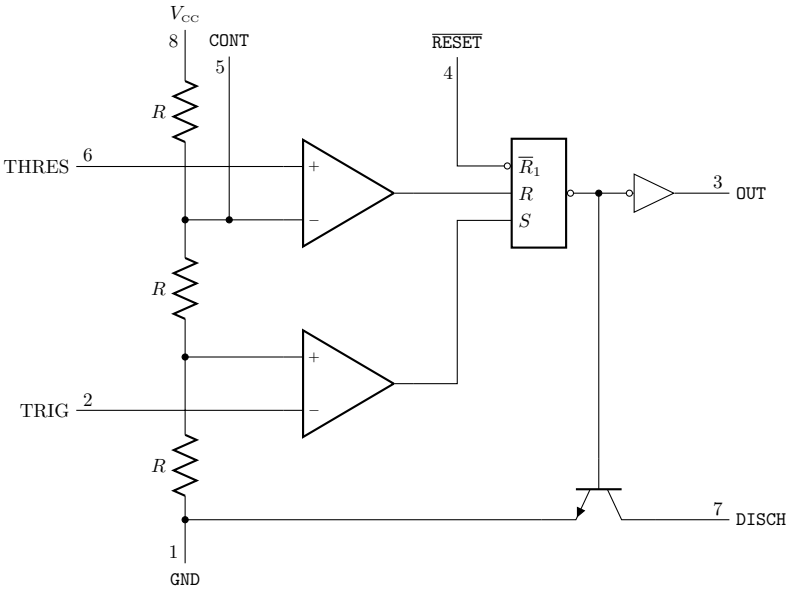


15.2 Monostable Multivibrators

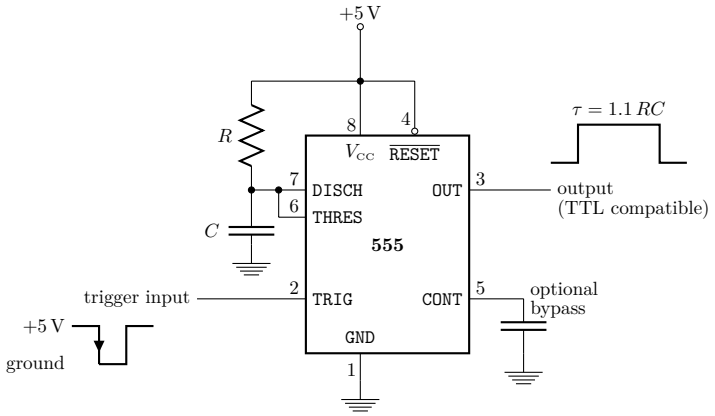
A **monostable multivibrator**, or **one-shot**, is a circuit that simply supplies a single digital pulse of some defined duration. This is the fundamental building block of timing circuits—circuits that control the timing of sequences of events (laser pulses, data acquisition, camera triggers, etc.).

15.2.1 555 as a One-Shot

The 555 can be conveniently hooked up as a one-shot. Recall the internals of the 555:



The connection as a one-shot is shown below. Again, the timing is set by the external components.



Let's see how this works:

1. The `TRIG` input is normally high, which is consistent with the internal flip-flop being reset, and thus the output being `LOW`. In this case the transistor is on, so the capacitor is shorted to ground. We will assume all this to be the case; if the flip-flop is indeed set, it is as if the circuit has been triggered, and the circuit will go through the rest of the cycle and reset, and then we can proceed with the assumption of a reset flip-flop.

2. The negative edge of a trigger pulse starts the output pulse. This happens when the `TRIG` input crosses below $(1/3)V_{CC}$, and the lower comparator sets the flip-flop. This takes the output `HIGH`, and turns the transistor off.

3. The timing of the output pulse is controlled by charging the capacitor C. The capacitor charges from V_{CC} via R.

4. The capacitor is uncharged at the beginning of the pulse, and it will charge until the capacitor voltage (and thus `THRES` input) rises to $(2/3)V_{CC}$. At this time, the upper comparator resets the flip-flop, ending the pulse and dumping the capacitor charge.

5. Again, we can use the exponential relaxation of the RC circuit to figure out the pulse duration. The capacitor voltage rises exponentially (with time RC) from 0 to $(2/3)V_{CC}$ towards V_{CC}. This is equivalent to an exponential decay from V_{CC} towards 0 to $(1/3)V_{CC}$. Thus we can consider the decay

$$V(t) = V_{CC}\, e^{-t/RC}, \tag{15.5}$$

and if we set $V(\tau) = (1/3)V_{CC}$, where τ is the pulse duration, we get

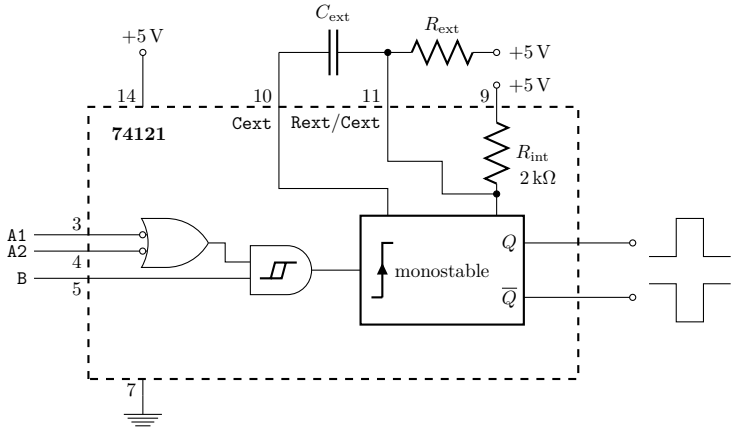
$$\tau = (\log 3)RC \approx 1.1\, RC. \tag{15.6}$$

Again, "log" here is the natural logarithm, and note that this is independent of the supply voltage.

Note that we have assumed that by the time the flip-flop is reset, the trigger pulse at the `TRIG` input is back to `HIGH`; otherwise the `S` input to the flip-flop will also be high, and we will put the flip-flop into the "bad" state. The net effect is that the output pulse will be "stretched" beyond the RC duration until the trigger input goes high. One workaround for this (if you're stuck with long pulses but want to trigger short pulses) is to run the trigger input through a differentiator (with a short RC time) and then into the 555.

15.2.1.1 The 74121

A number of other monostable multivibrators are available, more-or-less prepackaged. A good example is the 74121, which is faster than the 555 circuit—it can be programmed for pulses down to 35 ns, and up to 28 s. The "guts" and some external connections of the 74121 are shown below.



Some operation notes:

1. As in the 555, the internal monostable is controlled by the time for the external capacitor to charge via a resistor. The resistor can be internal (R_{int}) if pin 9 is connected to $+5\,V$, or external, if R_{ext} is connected between pin 11 and $+5\,V$. (Of course, if both are connected, then we have the parallel resistance of the two.) The pulse duration is given by

$$\tau = (\log 2)RC \approx 0.693\, RC. \tag{15.7}$$

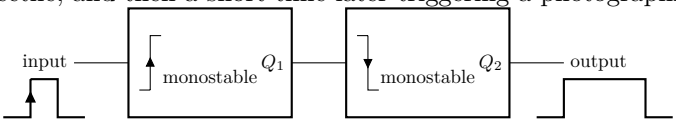
2. The internal monostable is triggered by a rising pulse edge, but there is some extra input logic to make this more flexible. For example, suppose `A1` and `B` are held `HIGH`. Then a falling edge on `A2` triggers the pulse. Of course, `A1` and `A2` can be exchanged here.

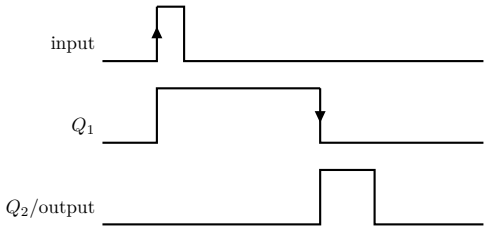
3. Similarly, if either `A1` or `A2` are held `LOW`, then a rising edge into `B` will trigger the flip-flop. In this case, the `B` input goes into a Schmitt trigger (0.2-V hysteresis), and thus can handle slow/noisy inputs.

4. The 74121 is **nonretriggerable**; this means the device will ignore any input edges while it is generating an output pulse. Other one-shots, like the 74123, is **retriggerable**, which means that a new triggering edge will always start a new timing cycle, even if already in the middle of a timing cycle.

15.2.1.2 Combining One-Shots: Pulse Delay

In complex timing systems, where many things must happen at the proper times, many one-shots can be chained together to generate the proper timing sequence. As simple example, two one-shots can be chained together to generate a *delayed* pulse from an initial trigger pulse. This times two events with a fixed delay, such as launching a projectile, and then a short time later triggering a photographic flash.



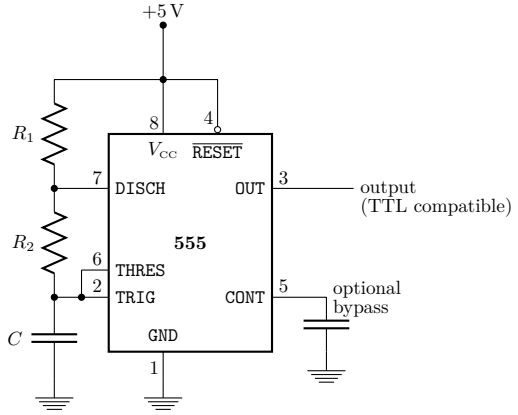


Note the different edge triggers of the two monostables.

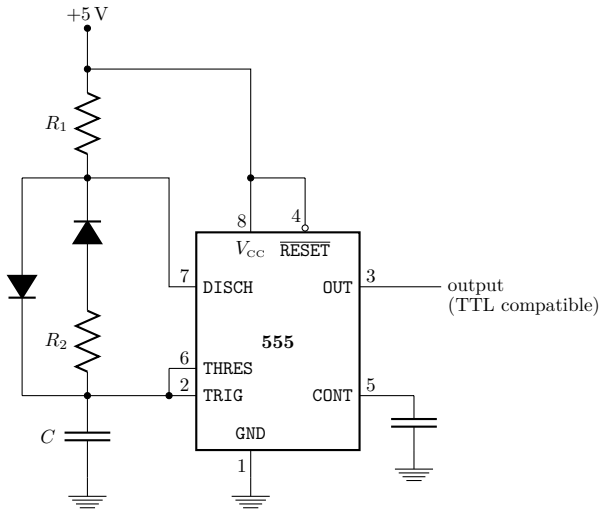
15.3 Circuit Practice

15.3.1 Duty-Cycle Control

What is the duty cycle of the basic astable circuit (reproduced below)?



What is the duty cycle of the modified astable shown below? (Ignore any voltage drops across the diodes.) How does it allow better control over the duty cycle?



Solution. With this arrangement, the capacitor charges through R_1 and the left-hand diode, and discharges through R_2 and the right-hand diode. In either case, we can work out the timing as follows. Suppose the capacitor discharges from voltage $(2/3)V$ towards zero, with time constant RC. We need to solve for the time τ when the capacitor voltage is $(1/3)V$:

$$\frac{2}{3}Ve^{-\tau/RC} = \frac{V}{3}. \tag{15.8}$$

The solution is

$$\tau = (\log 2)RC. \tag{15.9}$$

The `HIGH` output cycle is the charging cycle, so the high time is

$$\tau_{\texttt{HIGH}} = (\log 2)R_1C. \tag{15.10}$$

The `LOW` output cycle is the discharge cycle, so the high time is

$$\tau_{\texttt{LOW}} = (\log 2)R_2C. \tag{15.11}$$

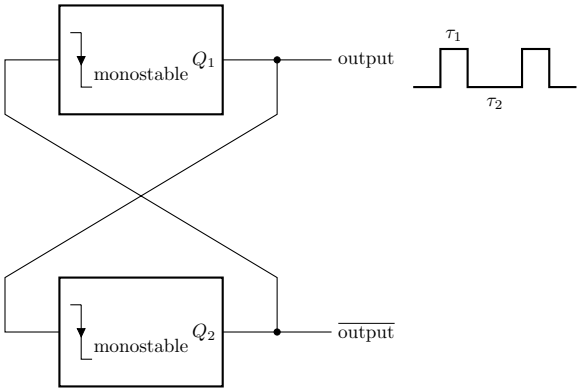
The period is the sum of these, or

$$T = (\log 2)(R_1 + R_2)C. \tag{15.12}$$

15.3.2 Astable Multivibrator

Another example of combining one-shots is to combine two, in order to make an astable multivibrator. How can you do this?

Solution. The circuit for this is shown below.

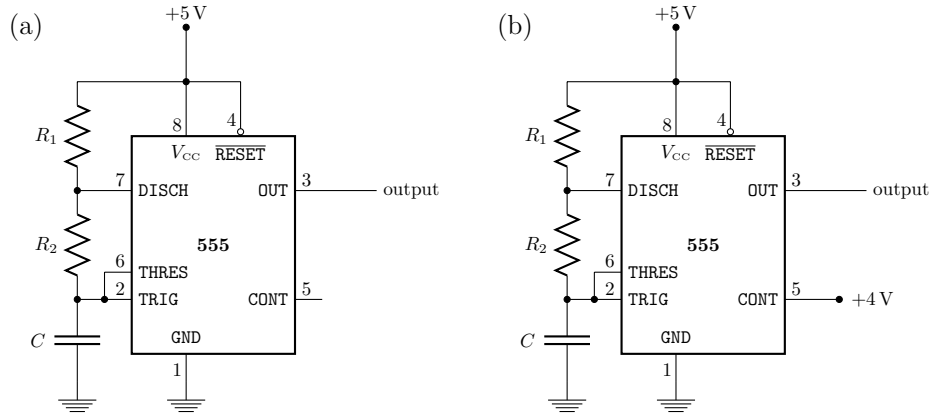


Note that the `HIGH` and `LOW` times are controlled separately by the durations of the two one-shots, and each one-shot triggers on the falling edge of the other.

15.4 Exercises

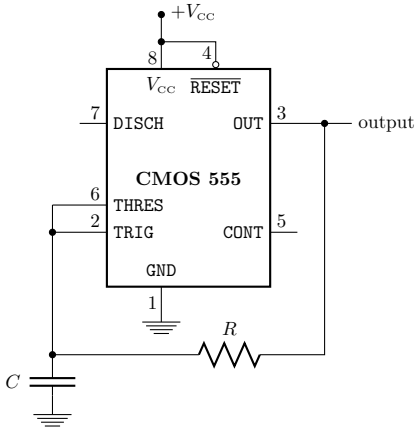
Problem 15.1

Shown below is (a) the basic 555 astable multivibrator and (b) a modified version. For each circuit compute the period and duty cycle in terms of R_1, R_2, C, and any relevant voltages.



Problem 15.2

Shown below is a 555-based astable multivibrator with 50% duty cycle. Show that it works as advertised, and compute the oscillation period.



Note: this circuit only works with CMOS variants of the 555 (like the 7555), where the output will swing all the way from ground to $+V_{CC}$.

Problem 15.3

Suppose you have a standard TTL logic gate, and you want to put a Schmitt trigger on one of the inputs to address a problematic input signal. However, suppose that all you have is a 555; show how to wire a 555 to act as a Schmitt trigger (actually, an inverting Schmitt trigger). Make sure to explain *why* your circuit acts as a Schmitt trigger, and what are the input logic levels.

Problem 15.4

Work out the period of the basic 555 astable multivibrator circuit, but this time account for the voltage drop V_d across the discharge transistor when it is turned on. That is, the supply voltage should no longer drop out of the result. Assuming $V_d = 0.2\,\text{V}$, how much does this affect the period compared to the idealized value if $V_{CC} = +5\,\text{V}$?

Problem 15.5

In one incarnation, the **bicolor LED** looks like an ordinary LED (plastic package with two leads), but is really 2 LEDs in parallel, with one reversed. That is, if current flows "forwards," the LED lights green, and if it flows "backwards," the LED lights red.

Design a circuit *two* 555's to light a bicolor LED, alternating between red and green. For concreteness, design for a ∼50% duty cycle (approximately equal time in each color), with a period of 1 s. Use whatever passive components you like, but be specific about their values. Also, show all pin connections on the 555's.

Hint: note that it would be a bad idea to configure two 555 as independent oscillators. Why? Instead try using one of the 555's as a `NOT` gate.

Problem 15.6

Design a 10 kHz square-wave oscillator (50% duty cycle) using only 74121's and capacitors.

Chapter 16

Digital–Analog Interfaces

One of the most important concepts in digital electronics is *interfacing* digital circuits to analog circuits. If an analog signal serves as the input to a digital circuit, then we need **analog-to-digital conversion (ADC)**, while a digital circuit generating an analog signal requires **digital-to-analog conversion (DAC)**. We will consider the latter first, which is simpler, and ADC often relies on DAC.

16.1 Digital-to-Analog Conversion

Digital-to-analog conversion is very common in everyday circuits. This is required to generate the audio signals in cell phones and CD/DVD/MP3 players, and to generate the output intensity (or color) of displays in LCD projectors or in CRT/plasma/LCD displays. Essentially, any analog signal coming out of a computer must have gone through the DAC process.

16.1.1 Resolution

Before understanding how DAC circuitry works, let's review some of the resolution requirements for representing analog signals. Analog signals must be **sampled**—that is instead of a continuous function $y(t)$, we must represent it via samples $y_j := y(t_j)$ at sample times t_j (typically regularly spaced), and the values of y_j must be represented with some finite precision (i.e., it must be represented with a finite number of bits). In terms of amplitude resolution, if there are N bits of data, then there are 2^N different signal levels available within a defined range (e.g., within some voltage range). The signal levels could be positive only, represented by unsigned integers, or positive/negative using signed integers (or unsigned integers after adding an offset that ensures the signal is always positive). In this case, since the "real" signal must always be rounded to the nearest available level, the *fractional* sampling resolution is 2^{-N}, and the absolute resolution is $2^{-N}V_{\text{range}}$ for a voltage signal if V_{range} is the total voltage range available for sampling the signal. Since the rounded value should be to the nearest sampling value, the maximum error is $1/2$ of the resolution, or $2^{-(N+1)}$ maximum fractional error for signals within V_{range}. So, for 16-bit sampling, the error is at worst about 8 ppm.

In terms of timing resolution, the requirement is set via the **sampling theorem**. Suppose we sample a signal every Δt in time. Then the **sampling rate** is given by

$$\text{sampling rate} = \frac{1}{\Delta t}. \tag{16.1}$$

Then we can also define the **Nyquist frequency** by

$$\text{Nyquist frequency} = \frac{1}{2\Delta t} = \frac{\text{sampling rate}}{2}. \tag{16.2}$$

It turns out that, according to the sampling theorem, the Nyquist frequency is the largest frequency that is accurately reproduced by the sampled signal.[1] For example, in compact-disc (CD) audio, the sampling rate

[1] For details, see Daniel Adam Steck, *Quantum and Atom Optics*, available online at `http://steck.us/teaching`.

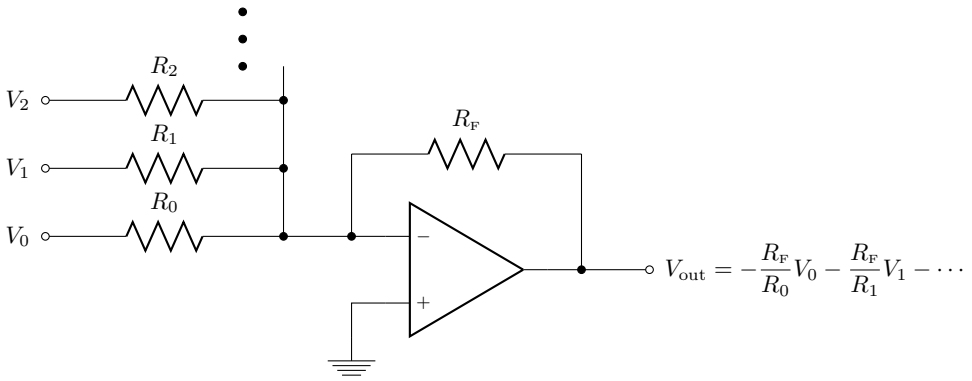
is 44.1 kHz, since the goal is to reproduce audio frequencies up to about 20 kHz. Note that 44.1 kHz is then a bit above the Nyquist frequency, which allows for extra tricks, like an **anti-aliasing filter**, to improve the quality of the reconstructed audio. The idea here is to guard against **aliasing**, which is the error suffered by frequencies *above* the Nyquist frequency—they are spuriously represented as lower (sub-Nyquist) frequencies in the sampled signal. Generally, a low-pass filter is used to remove these high frequencies, but since the filter does not have a perfectly sharp cutoff, the extra sampling rate above the Nyquist frequency accommodates the desired audio range, while giving some bandwidth for the low-pass filter to have a significant effect before aliasing errors occur.

16.1.2 DAC Circuitry

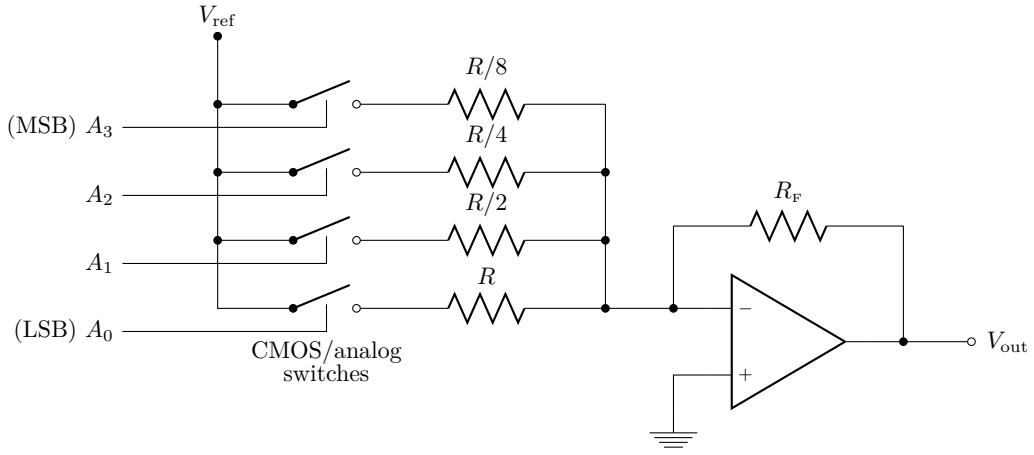
Now, how do we make a DAC? The basic ingredient is a summing (inverting) amplifier, which you may recall from Section 7.3.4. Recall that this takes a number of voltage inputs V_0, V_1, ..., and has as output

$$V_{\text{out}} = -\frac{R_{\text{F}}}{R_0}V_0 - \frac{R_{\text{F}}}{R_1}V_1 - \frac{R_{\text{F}}}{R_2}V_2 - \cdots, \tag{16.3}$$

which is an inverted, weighted sum, where the relative weights are controlled by the input resistors, and the weights have the feedback resistance R_{F} in common.



Then, for example, we can build a 4-bit DAC as in the diagram below.



Here, A_0–A_3 are digital inputs, with A_0 the LSB and A_3 the MSB of an unsigned integer. The inputs drive analog switches, which conduct when $A_j = \texttt{1}$ and are open when $A_j = \texttt{0}$. The voltage V_{ref} sets the (absolute)

voltage resolution and the range of the conversion. Then using the summing-amp formula,

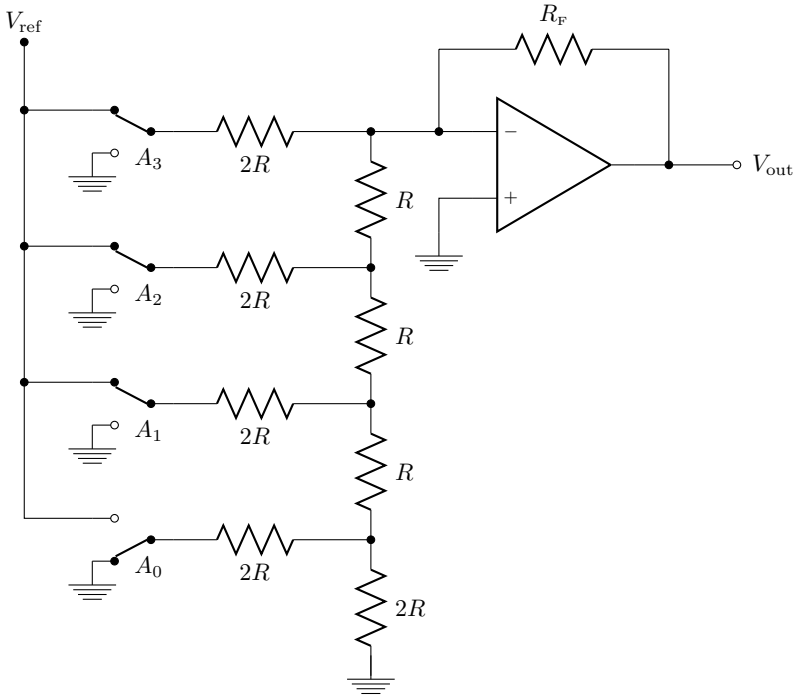
$$\begin{aligned} V_{\text{out}} &= -A_0\frac{R_{\text{F}}}{R}V_{\text{ref}} - 2A_1\frac{R_{\text{F}}}{R}V_{\text{ref}} - 4A_2\frac{R_{\text{F}}}{R}V_{\text{ref}} - 8A_3\frac{R_{\text{F}}}{R}V_{\text{ref}} \\ &= -\frac{R_{\text{F}}}{R}V_{\text{ref}}\left(A_0 2^0 + A_1 2^1 + A_2 2^2 + A_3 2^3\right). \end{aligned} \tag{16.4}$$

(DAC output)

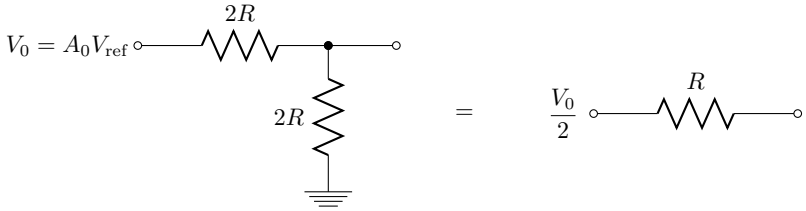
So, for example, an input of `0011` corresponds to $V_{\text{out}} = -3(R_{\text{F}}/R)V_{\text{ref}}$. Of course, usually we want a *positive* output, which requires another inverting amplifier.

16.1.3 R–2R Ladder

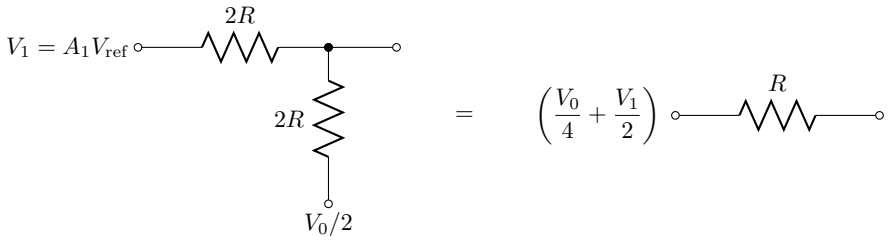
One problem with the above circuit is that it requires a series of resistors of different values to be made to high accuracy, which is difficult, especially for a high-resolution DAC. It is much easier to make sets of *matching* resistors, and there is a circuit that takes advantage of this, called the **R–2R ladder**, which uses only resistors of size R and $2R$. The idea is below.



Again, the inputs are A_0–A_3 here, controlling analog/CMOS switches (here SPDT). To see how this works, consider the Thévenin-equivalent circuit for the A_0 input. This is a simple 50% voltage divider, so the equivalent resistance is R, and the voltage is half the input $A_0 V_{\text{ref}}$. (Note that we consider $A_0 =$ `0` if the switch is down, $A_0 =$ `1` if the switch is up.)



Now lumping this equivalent circuit into the next "stage" with the A_1 input, we have a similar voltage-divider situation.



Note that the Thévenin-equivalent voltage is just the average of the two input voltages. Continuing this process, we find that at each stage we add in half of the next input voltage and divide the remaining ones by two. The result is

$$\begin{aligned} V_{\text{out}} &= -V_{\text{ref}} \frac{R_{\text{F}}}{R} \left(\frac{A_0}{16} + \frac{A_1}{4} + \frac{A_2}{8} + \frac{A_3}{2} \right) \\ &= -\frac{V_{\text{ref}} R_{\text{F}}}{2^4 R} \left(A_0 2^0 + A_1 2^1 + A_2 2^2 + A_3 2^3 \right). \end{aligned} \tag{16.5}$$

(output of R–$2R$ ladder)

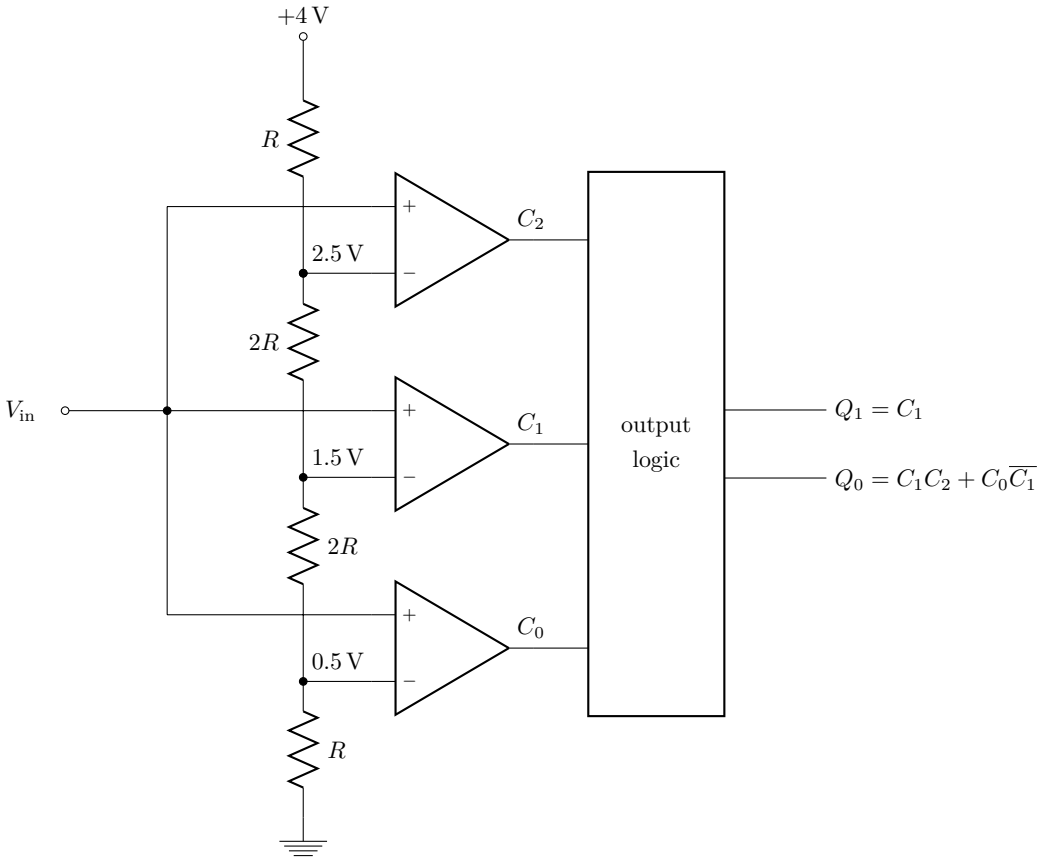
Thus, up to an overall factor, we obtain the same output as in Eq. (16.4).

16.2 Analog-to-Digital Conversion

The complementary process to digital-to-analog conversion is **analog-to-digital conversion (ADC)**. We will go through several ADC methods.

16.2.1 Flash ADC

A conceptually simple method for ADC is **flash ADC** or **parallel-encoding ADC**. The idea is to use a voltage-divider chain to create many reference voltage, and a separate comparator is used to compare the input voltage to each reference. Then output logic is needed to properly encode the digital output as a binary number. As a simple example, consider the 2-bit flash ADC below.



The idea is to choose digital voltage levels ("quantization level") of 0, 1, 2, and 3 V. Then the maximum input range is −0.5–3.5 V with a maximum error of 0.5 V. Then the conversion ranges with comparator outputs are enumerated below.

voltage range	digital output	comparator output $C_2C_1C_0$
−0.5–0.5 V	00	000
0.5–1.5 V	01	001
1.5–2.5 V	10	011
2.5–3.5 V	11	111

Note the logical expressions included to encode the three comparator outputs into two-bit binary.

The main advantage of a flash ADC is that it is *fast*: the signal just needs to propagate through the comparators and gates, and the ADC can sample rapidly changing signals. The main disadvantage is that for N bits, there must be 2^{N-1} comparators, which is difficult for more than about 10 bits of resolution.

16.2.2 Successive Approximation

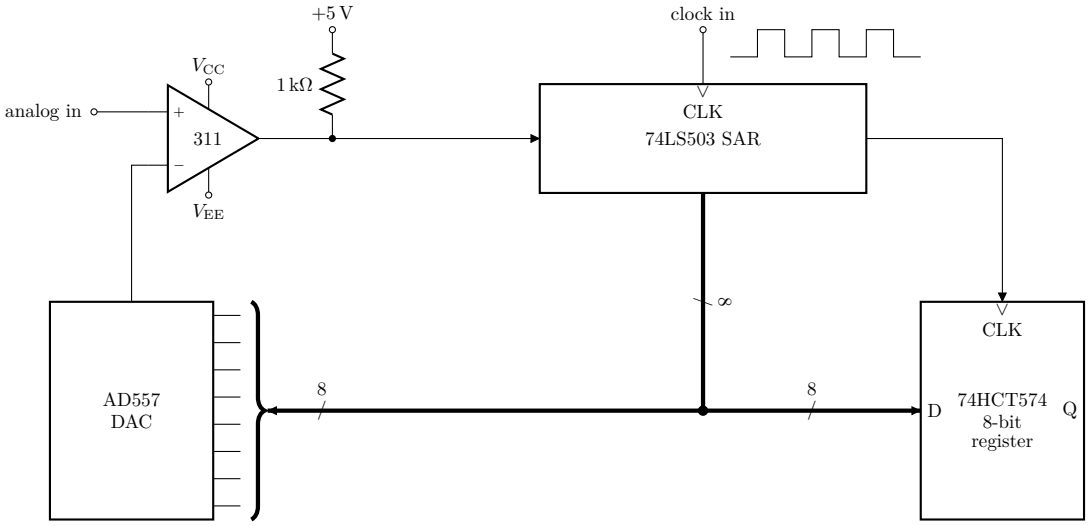
A slower, but more generally useful ADC method is **successive approximation**. This method is analogous to the root-finding problem: Suppose $f(x)$ is a continuous function with a single root in (a, b). That is, $f(a)f(b) < 0$. Then how do we find the root; i.e., how do we find x_0 such that $f(x_0) = 0$? The **bisection method** for root-finding works as follows.

1. We know (a, b) brackets x_0. So let $\bar{x}_0 = (a + b)/2$ be the initial best estimate for x_0.
2. If $f(a)f(\bar{x}_0) < 0$, then $(a, \bar{x}_0)$ brackets the root.
3. Otherwise, $(\bar{x}_0, b)$ brackets the root.

4. Redefine (a, b) to be the new, tighter bracketing interval for x_0, and repeat.

This process converges exponentially, because the width of the bracketing interval is halved on each iteration. If $\Delta x = b - a$, then the initial worst-case error in the estimate $\bar{x}_0$ is $\Delta x/2$. After N iterations, the error is $\Delta x/2^{N+1}$.

The successive-approximation approach to ADC is the same problem, but to find what digital voltage $\bar{V}$ best corresponds to V_{in}. This is an iterative process to a predetermined accuracy, given by the number of digital bits. A circuit to implement this procedure is shown below. This uses the 74LS503 **successive-approximation register (SAR)** (now obsolete), which controls the bisection process. Most ADCs nowadays using successive approximation have all these components integrated into a single chip, with serial data output, which is handy for keeping pin counts low, but makes it harder to understand what is going on inside.



The SAR, the "mastermind" of the conversion process, works to find 1 bit of the digital result on each clock cycle, starting with the MSB. It does this by writing the "midpoint value" of the DAC's range to the DAC (as the first approximation), and reads the comparison result from the comparator, which tells the SAR which half of the DAC range brackets V_{in}. On the next cycle, the SAR writes out the new midpoint of the smaller bracketing range, and records the comparison result as the next converted bit, and so on. The 8-bit latch ('574) holds the completed conversion, while the SAR is performing the next conversion (and thus its outputs are changing).

As a 2-bit example, consider the same analog range (digital levels of 0, 1, 2, and 3 V). But now, we will consider the conversion ranges to be

$$\begin{aligned} &< 0\,\text{V} = \texttt{00} \\ &0\text{–}1\,\text{V} = \texttt{01} \\ &1\text{–}2\,\text{V} = \texttt{10} \\ &> 2\,\text{V} = \texttt{11}, \end{aligned} \tag{16.6}$$

as we will see. Note the different offset compared to the flash-ADC example. Let's assume a 1.3-V input. Then the process is as follows:

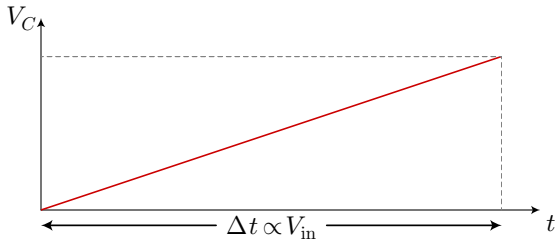
- During clock cycle #1, the SAR tries the midpoint of the whole range. We can take this to be `01` (more generally, `0111111...` for N bits). The DAC voltage is then 1 V, so the comparator is `HIGH`, and so the MSB is `1`.
- During clock cycle #2, the SAR tries the midpoint of the remaining range (`10`–`11`). The "midpoint" is `10`. The DAC voltage is then 2 V, so the comparator is now `LOW`, and so the LSB is `0`.

Thus, the converted result is `10`, in agreement with the table above.

The advantages of SA-ADC is that the timing is guaranteed (measured in clock cycles), the result can be very accurate (if a good DAC is used), and the circuit is not too complicated. The main disadvantage are that SA-ADC is slower than flash conversion, and thus may need a sample/hold circuit to deal with rapidly changing input signals.

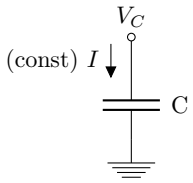
16.2.3 Single/Dual-Slope ADC

Another pair of important ADC methods goes under the name(s) of **single/dual-slope ADC**. The basic idea in **single-slope ADC** is to use a constant-current source to charge a capacitor, and then use a counter/clock combination to measure the time for the capacitor voltage to reach the input V_{in}.



The charge time is then proportional to the voltage. The idea is that time is very easy to measure accurately and precisely, so the method can be very accurate.

To show this mathematically, consider the capacitor-charging situation diagramed below, where a constant-current source I charges a capacitor to voltage $V_C(t)$.



Then using $Q = CV_C$ and differentiating,

$$I = \frac{dQ}{dt} = C\frac{dV_C}{dt}, \tag{16.7}$$

so that for constant I,

$$V_C(t) = \frac{It}{C}. \tag{16.8}$$

Since I/C is a constant, this can be calibrated precisely to yield V_C (and thus V_{in}) in terms of t.

In **dual-slope ADC**, the conversion is done in two steps.

1. C is charged for a fixed time τ by a constant current $I \propto V_{\text{in}}$. If we let α be constant, then we can write $I = \alpha V_{\text{in}}$, so that
$$V_C(\tau) = \frac{\alpha V_{\text{in}}\tau}{C}. \tag{16.9}$$

2. Then, C is discharged at a constant current I', and the discharge time δt is measured. Then the discharge time is fixed as in single-slope ADC by
$$V_C(\tau) = \frac{I'\delta t}{C}. \tag{16.10}$$
Thus,
$$\delta t = \frac{CV_C(\tau)}{I'} = \frac{C(\alpha V_{\text{in}}\tau/C)}{I'} = \left(\frac{\alpha\tau}{I'}\right) V_{\text{in}}. \tag{16.11}$$

In this way, we are left to calibrate $\alpha\tau/I'$, which is a combination of the output of a current source (α), a time τ, and a current I', all of which can be well-calibrated. Notably, the capacitance C dropped out; capacitances are difficult to fabricate in a way that is accurate and stable. Also, note that the first stage takes time τ, which has an averaging effect over noise in V_{in}, whereas single-slope conversion is more apt to trigger early on a downward noise fluctuation of V_{in}.

16.3 Circuit Practice

Suppose you have a 3-bit DAC, with voltage levels 0 V, 0.1 V, 0.2 V, ..., in a successive-approximation ADC. If $V_{\text{in}} = 0.35\,\text{V}$,

- make a plot of the DAC output vs. time
- what is the final, converted digital value?

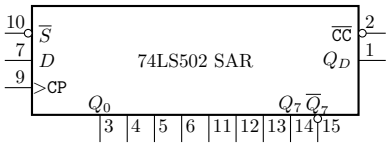
Answer. For the converted digital value: `100`. The comparison voltages will be: 0.3 V, 0.5 V, 0.4 V.

16.4 State-Machine Emulation of SARs

A nice example of an application of a general state machine is the emulation of classic, but obsolete, integrated circuits that acts as successive-approximation registers (SARs) for ADC circuits. We will discuss the application of two SARs, the 74LS502 and the 74LS503.[2]

16.4.1 74502

The 74LS502 is a 8-bit successive-approximation register (SAR) that is shown schematically below.



It operates as follows. There are three inputs:

- D is the input data (i.e., the output of the comparator feeds this input).
- `CP` is the clock-pulse input; the SAR action occurs on the rising pulse edges.
- $\overline{S}$ is the "start" input. This is normally `HIGH`, and brought `LOW` for one cycle (i.e., during one rising edge of `CP`) to begin conversion.

There are also 11 outputs:

- Q_0–Q_7 are the (parallel) digital conversion outputs.
- $\overline{Q_7}$ is an inverted copy of Q_7 (useful for signed-value ADC).
- $\overline{CC}$ is `HIGH` during conversion, and goes `LOW` when the ADC operation is complete.
- Q_D is a synchronized copy of D (i.e., Q_D latches the state of D at the last rising clock edge), which could be used for serial data output.

Pin 8 ground and pin 16 power are not shown in the schematic diagram.

The operational rules are as follows.

[2]Related files and more information available at `http://atomoptics.uoregon.edu/~dsteck/74503`.

- On a LOW start pulse on $\overline{S}$, the chip sets $Q_7Q_6\cdots Q_0 = 01111111$, and $\overline{CC} = 1$.
- Next clock pulse: set $Q_7 = D$, $Q_6 = 0$, $Q_D = D$.
- Next clock pulse: set $Q_6 = D$, $Q_5 = 0$, $Q_D = D$

  ⋮
- Next clock pulse: set $Q_1 = D$, $Q_0 = 0$, $Q_D = D$.
- Next clock pulse: set $Q_0 = D$, $\overline{CC} = 0$.
- On subsequent clock pulses, we only care that the parallel data Q_0–Q_7 and $\overline{CC}$ do not change.

We can also summarize this via the truth table (X = "don't care"):

clock	inputs		outputs									
cycle	D	$\overline{S}$	Q_D	Q_7	Q_6	Q_5	Q_4	Q_3	Q_2	Q_1	Q_0	$\overline{CC}$
0	X	0	X	X	X	X	X	X	X	X	X	X
1	D_7	1	X	0	1	1	1	1	1	1	1	1
2	D_6	1	D_7	D_7	0	1	1	1	1	1	1	1
3	D_5	1	D_6	D_7	D_6	0	1	1	1	1	1	1
4	D_4	1	D_5	D_7	D_6	D_5	0	1	1	1	1	1
5	D_3	1	D_4	D_7	D_6	D_5	D_4	0	1	1	1	1
6	D_2	1	D_3	D_7	D_6	D_5	D_4	D_3	0	1	1	1
7	D_1	1	D_2	D_7	D_6	D_5	D_4	D_3	D_2	0	1	1
8	D_0	1	D_1	D_7	D_6	D_5	D_4	D_3	D_2	D_1	0	1
9	X	1	D_0	D_7	D_6	D_5	D_4	D_3	D_2	D_1	D_0	0
10	X	1	X	D_7	D_6	D_5	D_4	D_3	D_2	D_1	D_0	0

16.4.1.1 General Emulation Notes

Now we will review the logic to implement this chip as a synchronous state machine. We will need (flip-flop) register outputs for Q_0–Q_7, Q_D, and $\overline{CC}$. We will also use three extra register bits C_2, C_1, and C_0, where $C_2C_1C_0$ gives (in binary) the next parallel-output bit to latch. That is, the circuit should set $C_2C_1C_0 = 111$ on clock cycle 1 in the truth table, and then count backwards to 000.

To get started, first note that $\overline{Q}_7$ can just be implemented with an extra NOT gate at the output of Q_7, and does not require its own register output. Next, note that we can implement

$$D_{Q_5} = \overline{\overline{S}} + (C_2\overline{C_1}C_0D + \overline{C_2\overline{C_1}C_0}Q_5)\overline{C_2C_1\overline{C_0}}. \tag{16.12}$$

Read this as follows. If the start pulse is 0 (so $S = \overline{\overline{S}} = 1$), then force $Q_5 = 1$, which is accomplished by the first term. The second term has an overall multiplier that forces the expression to 0 (unless there is an override by $\overline{S}$) when the $C_2C_1C_0$ is at 6. The D term then stores the input data D when $C_2C_1C_0$ is at 5, and the Q_5 term allows Q_5 to persist for other values of $C_2C_1C_0$.

As an example of a counter bit, note that C_1 should change when $C_2C_1C_0$ is X10 or 100 (remember we are counting backwards in binary). In either case it simply toggles, so

$$D_{C_1} = \overline{\overline{S}} + \overline{C_1\overline{C_0}}C_1 + C_2\overline{C_1}\,\overline{C_0}. \tag{16.13}$$

The first term forces this bit to 1 on the start pulse, the last term forces the bit to 1 on $C_2C_1C_0 = 100$, and the middle term is zero on $C_2C_1C_0 = X10$, and allows C_1 to persist otherwise.

The complete set of logic expressions is as follows.

$$
\begin{aligned}
D_{Q_7} &= \overline{S}(C_2C_1C_0D + \overline{C_2C_1C_0}Q_7) \\
D_{Q_6} &= \overline{\overline{S}} + (C_2C_1\overline{C_0}D + \overline{C_2C_1\overline{C_0}}Q_6)\overline{C_2C_1C_0} \\
D_{Q_5} &= \overline{\overline{S}} + (C_2\overline{C_1}C_0D + \overline{C_2\overline{C_1}C_0}Q_5)\overline{C_2C_1\overline{C_0}} \\
D_{Q_4} &= \overline{\overline{S}} + (C_2\overline{C_1}\,\overline{C_0}D + \overline{C_2\overline{C_1}\,\overline{C_0}}Q_4)\overline{C_2\overline{C_1}C_0} \\
D_{Q_3} &= \overline{\overline{S}} + (\overline{C_2}C_1C_0D + \overline{\overline{C_2}C_1C_0}Q_3)\overline{C_2\overline{C_1}\,\overline{C_0}} \\
D_{Q_2} &= \overline{\overline{S}} + (\overline{C_2}C_1\overline{C_0}D + \overline{\overline{C_2}C_1\overline{C_0}}Q_2)\overline{\overline{C_2}C_1C_0} \\
D_{Q_1} &= \overline{\overline{S}} + (\overline{C_2}\,\overline{C_1}C_0D + \overline{\overline{C_2}\,\overline{C_1}C_0}Q_1)\overline{\overline{C_2}C_1\overline{C_0}} \\
D_{Q_0} &= \overline{\overline{S}} + (\overline{C_2}\,\overline{C_1}\,\overline{C_0}\,\overline{\texttt{CC}}D + \overline{\overline{C_2}\,\overline{C_1}\,\overline{C_0}\,\overline{\texttt{CC}}}Q_0)\overline{\overline{C_2}\,\overline{C_1}C_0} \\
D_{\overline{\texttt{CC}}} &= \overline{\overline{S}} + \overline{\overline{C_2}\,\overline{C_1}\,\overline{C_0}\,\overline{\texttt{CC}}} \\
D_{C_2} &= \overline{\overline{S}} + \overline{C_2\overline{C_1}\,\overline{C_0}}\,C_2 \\
D_{C_1} &= \overline{\overline{S}} + \overline{C_1\overline{C_0}}C_1 + C_2\overline{C_1}\,\overline{C_0} \\
D_{C_0} &= \overline{\overline{S}} + \overline{\overline{C_2}\,\overline{C_1}\,\overline{C_0}}\,\overline{C_0} \\
D_{Q_D} &= D \\
\overline{Q}_7 &= \overline{(Q_7)}
\end{aligned}
\tag{16.14}
$$

To understand all these in some detail:

- The easiest ones to understand are the last two: $\overline{Q}_7$ is just an inverted copy of Q_7, and Q_D just latches D.
- On all the others, the start pulse forces the bits to `1`, except for Q_7, which is forced to `0`. Note that the start pulse overrides all other information.
- Then Q_7 loads D when $C_2C_1C_0 = \texttt{111}$ and persists otherwise. We covered the similar logic for Q_5 already; Q_0–Q_4 and Q_6 follow the same idea in persisting, except changing to `0` and storing D at the right stages. However, Q_0 is slightly more complicated since we stop the counter on $C_2C_1C_0 = \texttt{000}$, so we must also use the $\overline{\texttt{CC}}$ bit to make sure the state of Q_0 persists when conversion is complete.
- For $\overline{\texttt{CC}}$, this should be forced to zero on $C_2C_1C_0 = \texttt{000}$ (i.e., after the last comparison), otherwise it persists. The factor of $\overline{\texttt{CC}}$ is not strictly necessary on the second term.
- We covered the logic of the C_1 counter bit already. Then C_0 is simple in toggling on each clock pulse, except that once $C_2C_1C_0 = \texttt{000}$, we will force it to stay at `0`. For C_2, the MSB should only change on $C_2C_1C_0 = \texttt{100}$, so we detect

16.4.1.2 22V10 Emulation

To emulate the 74LS502, we will choose the 22V10 SPLD (e.g., the ATF22V10C from Atmel). This has a 10-bit register (i.e., 10 D-type flip-flops) and 10 outputs, plus plenty of logic for sum-of-product logic and plenty of inputs. However, there is a problem: we have 13 registered outputs, but only 10 bits in the register. To handle this, note that the counter-logic bits C_0–C_2 are actually redundant–and-that is, although they are conceptually useful states in writing down the state-machine logic, they are not needed as register variables in the sense that they can be inferred from the other register variables Q_0–Q_7 and $\overline{\texttt{CC}}$. In particular, note that

- $C_2C_1C_0 = \texttt{111}$ is determined by $Q_7 \cdots Q_0\overline{\texttt{CC}} = \texttt{011111111}$
- $C_2C_1C_0 = \texttt{110}$ is determined by $Q_7 \cdots Q_0\overline{\texttt{CC}} = \texttt{X01111111}$
- $C_2C_1C_0 = \texttt{101}$ is determined by $Q_7 \cdots Q_0\overline{\texttt{CC}} = \texttt{XX0111111}$
- etc.

Thus, C_0 should trigger on an even number of ones after the leading zero:

$$\begin{aligned} C_0 &= \overline{Q_7}Q_6Q_5Q_4Q_3Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_5}Q_4Q_3Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_3}Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_1}Q_0\overline{\texttt{CC}} \\ &= \big(\big(\big(\overline{Q_7}Q_6Q_5 + \overline{Q_5}\big)Q_4Q_3 + \overline{Q_3}\big)Q_2Q_1 + \overline{Q_1}\big)Q_0\overline{\texttt{CC}} \\ &= \big(\big(\big(\overline{Q_7}Q_6 + \overline{Q_5}\big)Q_4 + \overline{Q_3}\big)Q_2 + \overline{Q_1}\big)Q_0\overline{\texttt{CC}}. \end{aligned} \tag{16.15}$$

We used $A + \overline{A}B = A + B$ in the last step. Similarly C_1 triggers when there are 8, 7, 4, or 3 ones after the leading zero:

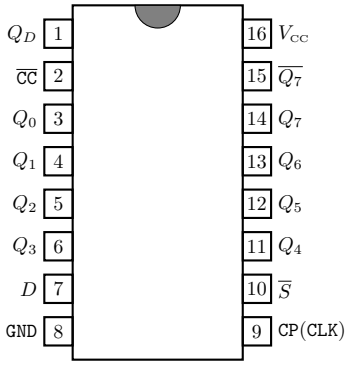
$$\begin{aligned} C_1 &= \overline{Q_7}Q_6Q_5Q_4Q_3Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_6}Q_5Q_4Q_3Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_3}Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_2}Q_1Q_0\overline{\texttt{CC}} \\ &= \big(\big(\big(\overline{Q_7}Q_6 + \overline{Q_6}\big)Q_5Q_4Q_3 + \overline{Q_3}\big)Q_2 + \overline{Q_2}\big)Q_1Q_0\overline{\texttt{CC}} \\ &= \big(\big(\overline{Q_7} + \overline{Q_6}\big)Q_5Q_4 + \overline{Q_3} + \overline{Q_2}\big)Q_1Q_0\overline{\texttt{CC}}. \end{aligned} \tag{16.16}$$

Finally, C_2 triggers when there are 8, 7, 6, or 5 ones after the leading zero:

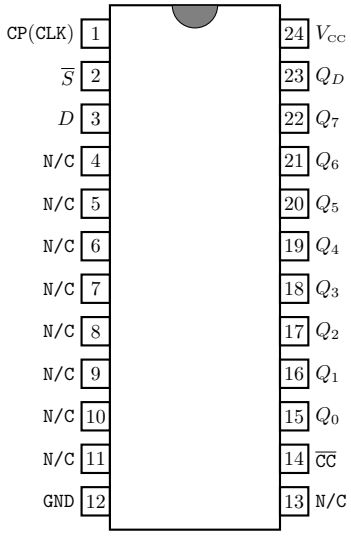
$$\begin{aligned} C_2 &= \overline{Q_7}Q_6Q_5Q_4Q_3Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_6}Q_5Q_4Q_3Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_5}Q_4Q_3Q_2Q_1Q_0\overline{\texttt{CC}} + \overline{Q_4}Q_3Q_2Q_1Q_0\overline{\texttt{CC}} \\ &= \big(\big(\big(\overline{Q_7}Q_6 + \overline{Q_6}\big)Q_5 + \overline{Q_5}\big)Q_4 + \overline{Q_4}\big)Q_3Q_2Q_1Q_0\overline{\texttt{CC}} \\ &= \big(\overline{Q_7} + \overline{Q_6} + \overline{Q_5} + \overline{Q_4}\big)Q_3Q_2Q_1Q_0\overline{\texttt{CC}}. \end{aligned} \tag{16.17}$$

Thus, we are down to 10 registered outputs, and we may proceed.

The pinout for the original 74LS502 and the 22V10-based emulator are shown below. Note that because we only have 10 outputs available, we have chosen to include the Q_D output but not the $\overline{Q_7}$ output, but we could easily make the opposite choice, as we will discuss below.



original 74LS502



74502 emulator

The code to implement the state machine in the **CUPL** (Compiler for Universal Programmable Logic) programming language[3] is shown below.

74502-22V10.pld

```
/*
*   74502 SAR emulator, on a 22v10
*/
Name            74502-22V10;
Partno          74502;
Revision        01;
Date            5/23/2015;
Designer        Daniel Steck;
Company         University of Oregon;
Location        None;
Assembly        None;
Device          g22v10;

/*** inputs ***/
pin 1 = CP;          /* clock pulse (trig on rising edge) */
pin 2 = !S;          /* start low  */
pin 3 = Din;         /* data */

/*** outputs ***/
pin 14       = !CC;       /* conversion complete low */
pin [15..22] = [Q0..Q7];  /* 8-bit output */
pin 23       = QD;        /* registered/synchronous copy of data input */

/*** intermediate counter variables ***/
C0 = (((!Q7 & Q6 # !Q5) & Q4 # !Q3) & Q2 # !Q1) & Q0 & !CC;
C1 = ((!Q7 # !Q6) & Q5 & Q4 # !Q3 # !Q2) & Q1 & Q0 & !CC;
C2 = (!Q7 # !Q6 # !Q5 # !Q4) & Q3 & Q2 & Q1 & Q0 & !CC;

/*** register inputs ***/
CC.D = !(S # !(!C2 & !C1 & !C0) & !CC);
Q7.D = !S & (C2 & C1 & C0 & Din # !(C2 & C1 & C0) & Q7);
Q6.D = S # ( C2 &  C1 & !C0 & Din # !( C2 &  C1 & !C0) & Q6) & !( C2 &  C1 &  C0);
Q5.D = S # ( C2 & !C1 &  C0 & Din # !( C2 & !C1 &  C0) & Q5) & !( C2 &  C1 & !C0);
Q4.D = S # ( C2 & !C1 & !C0 & Din # !( C2 & !C1 & !C0) & Q4) & !( C2 & !C1 &  C0);
Q3.D = S # (!C2 &  C1 &  C0 & Din # !(!C2 &  C1 &  C0) & Q3) & !( C2 & !C1 & !C0);
Q2.D = S # (!C2 &  C1 & !C0 & Din # !(!C2 &  C1 & !C0) & Q2) & !(!C2 &  C1 &  C0);
Q1.D = S # (!C2 & !C1 &  C0 & Din # !(!C2 & !C1 &  C0) & Q1) & !(!C2 &  C1 & !C0);
Q0.D = S # (!C2 & !C1 & !C0 & !CC & Din # !(!C2 & !C1 & !C0 & !CC) & Q0) & !(!C2 & !C1 &  C0);
QD.D = Din;

/*** handle flip-flop variables set/preset inputs ***/
CC.ar = 'b'0;
Q7.ar = 'b'0;
Q6.ar = 'b'0;
Q5.ar = 'b'0;
Q4.ar = 'b'0;
Q3.ar = 'b'0;
Q2.ar = 'b'0;
Q1.ar = 'b'0;
Q0.ar = 'b'0;
QD.ar = 'b'0;
```

[3]The only compiler realistically available nowadays is WinCUPL, which is Windows-based, crash-prone, and proprietary, but it is freely distributed by Atmel: `http://www.atmel.com/tools/wincupl.aspx`.

```
CC.sp = 'b'0;
Q7.sp = 'b'0;
Q6.sp = 'b'0;
Q5.sp = 'b'0;
Q4.sp = 'b'0;
Q3.sp = 'b'0;
Q2.sp = 'b'0;
Q1.sp = 'b'0;
Q0.sp = 'b'0;
QD.sp = 'b'0;
```

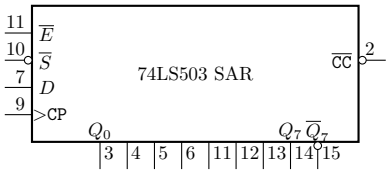
This is a relatively straightforward translation of the equations we have already written down. Note that

- The first block contains obligatory header information, most of which is merely informational, but the "Device" declaration to "g22v10" means we have selected this device (which also covers the ATF22V10C variant).
- The next two blocks declare input and output pin assignments and variables. Note that a `NOT` is denoted by "!"; for example, $\overline{S}$ is denoted `!S`.
- The next block gives the expressions (16.15)–(16.17) for the counter variables C_0–C_2. Note that the `OR` operation is represented by a hash (`#`), and the `AND` operation is represented by an ampersand (`&`).
- The next block gives expressions for all the register inputs, as in Eqs. (16.14) (except for counter register variables). The CUPL notation is that the input for the register variable `Q7` is `Q7.D` (i.e., this is what we call D_{Q_7}).
- Finally, in the last block, we make sure to tie the other flip-flop controls to default values. Here, the 22V10 flip-flops have asynchronous-reset (`AR`) and synchronous-preset (`SP`) inputs; we simply tie all of them to logical `0` (written as "binary `0`" or `'b'0` in CUPL).
- To have $\overline{Q_7}$ as an output instead of D_7, we can change the "`pin 23 = QD;`" declaration to now read "`pin 23 = notQ7;`", and then change the line "`QD.D = Din;`" to "`notQ7 = !Q7;`" (note that this output would then no longer be registered, but "combinatorial").

The compiler then simplifies the logical expressions and figures out how to make the fused connections in the configurable-logic section of the 22V10. A separate programmer is necessary to then "burn" the chip.

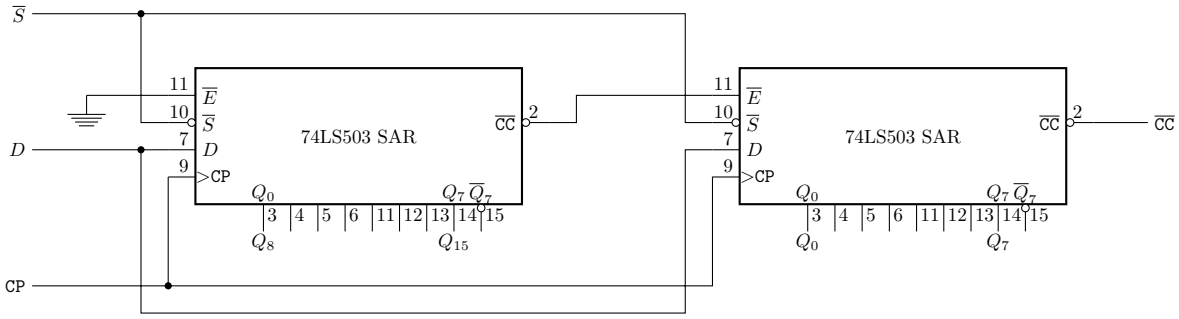
16.4.2 74503

The 74LS503 is basically the same as the '502, but it dispenses with Q_D and adds an enable-`LOW` input $\overline{E}$.



The $\overline{E}$ operates as follows. If it is held `LOW`, then the chip behavior is essentially identical to the '502. If it is taken `HIGH`—the intent is for this to happen after the start operation but before any data acquisition occurs—then Q_7 is *asynchronously* forced `HIGH`, and the chip does not accept any data from D. When $\overline{E} =$ `0` again, the acquisition process proceeds as in the '502.

The idea behind the $\overline{E}$ input in the '503 is that two '503's can be "stacked" to realize a 16-bit SAR. The idea is that the $\overline{CC}$ of the most-significant chip (byte) drives the $\overline{E}$ of the least-significant chip, so that when the first chip is finished, acquisition continues on the second chip. The connections are shown in the data sheet for the 74LS503. The idea is to share the data, clock, and start lines, and chain the $\overline{CC}$ of the MS chip to the enable of the enable of the LS chip.



16.4.2.1 ATF750C Emulation

To emulate the 74LS503, we will choose the ATF750C CPLD (from Atmel). Note that when the C_j are eliminated as in Eqs. (16.15)–(16.17) in the case of the '503, the expressions for D_{Q_0}–D_{Q_6} must have "$+E$" tacked on to each expression, since these variables track the counting state. However, this extra addition appears to make the logic too complicated to fit in either the 22V10 or the ATF750C.

Fortunately, the ATF750C has 10 extra register bits that are present "internally" (i.e., they can not be connected directly to outputs as are the other 10 register bits). Thus we can implement C_0–C_2 as register variables.

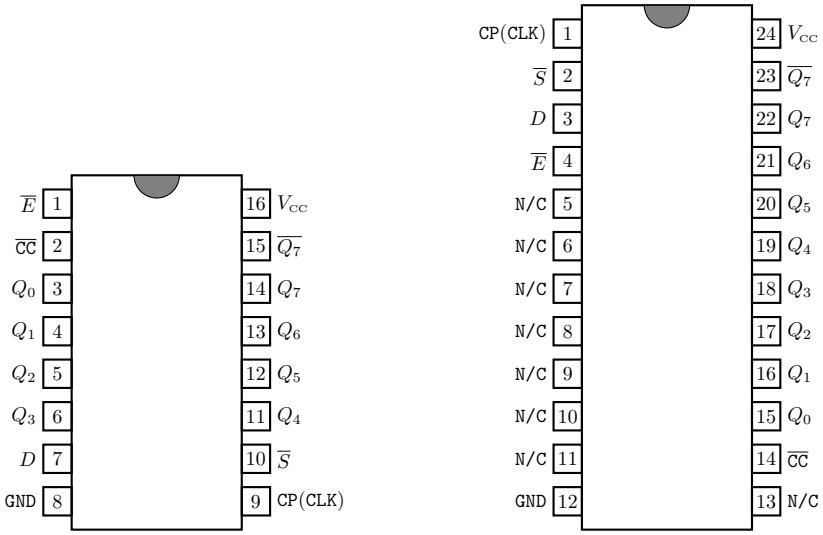
Now we need to show how to modify the '502 logic to accommodate this new input. Note that since the effect on Q_7 is asynchronous, Q_7 can't any more be a register output. So for the sake of notation, let P_7 be a register output, and let Q_7 be a Boolean function of P_7 and $\overline{E}$. The other outputs Q_0–Q_6 and $\overline{\texttt{CC}}$ can still be register outputs. The summary of logic expressions is below.

$$
\begin{aligned}
D_{P_7} &= \overline{S}(C_2C_1C_0D + \overline{C_2C_1C_0}P_7)\overline{\overline{E}} \\
Q_7 &= \overline{E} + P_7 \\
D_{Q_6} &= \overline{\overline{S}} + (C_2C_1\overline{C_0}D + \overline{C_2C_1\overline{C_0}}Q_6)\overline{C_2C_1C_0} \\
D_{Q_5} &= \overline{\overline{S}} + (C_2\overline{C_1}C_0D + \overline{C_2\overline{C_1}C_0}Q_5)\overline{C_2C_1\overline{C_0}} \\
D_{Q_4} &= \overline{\overline{S}} + (C_2\overline{C_1}\,\overline{C_0}D + \overline{C_2\overline{C_1}\,\overline{C_0}}Q_4)\overline{C_2\overline{C_1}C_0} \\
D_{Q_3} &= \overline{\overline{S}} + (\overline{C_2}C_1C_0D + \overline{\overline{C_2}C_1C_0}Q_3)\overline{C_2\overline{C_1}\,\overline{C_0}} \\
D_{Q_2} &= \overline{\overline{S}} + (\overline{C_2}C_1\overline{C_0}D + \overline{\overline{C_2}C_1\overline{C_0}}Q_2)\overline{\overline{C_2}C_1C_0} \\
D_{Q_1} &= \overline{\overline{S}} + (\overline{C_2}\,\overline{C_1}C_0D + \overline{\overline{C_2}\,\overline{C_1}C_0}Q_1)\overline{\overline{C_2}C_1\overline{C_0}} \\
D_{Q_0} &= \overline{\overline{S}} + (\overline{C_2}\,\overline{C_1}\,\overline{C_0}\,\overline{\texttt{CC}}D + \overline{\overline{C_2}\,\overline{C_1}\,\overline{C_0}\,\overline{\texttt{CC}}}Q_0)\overline{\overline{C_2}\,\overline{C_1}C_0} \\
D_{\overline{\texttt{CC}}} &= \overline{\overline{S}} + \overline{\overline{C_2}\,\overline{C_1}\,\overline{C_0}}\,\overline{\texttt{CC}} \\
D_{C_2} &= \overline{\overline{S}} + \overline{C_2\overline{C_1}\,\overline{C_0}}\,C_2 + \overline{E} \\
D_{C_1} &= \overline{\overline{S}} + \overline{C_1\overline{C_0}}C_1 + C_2\overline{C_1}\,\overline{C_0} + \overline{E} \\
D_{C_0} &= \overline{\overline{S}} + \overline{\overline{C_2}\,\overline{C_1}\,\overline{C_0}}\,\overline{C_0} + \overline{E} \\
\overline{Q}_7 &= \overline{(Q_7)}
\end{aligned}
\tag{16.18}
$$

Note that we are forcing $P_7 = \texttt{0}$ on a disable cycle, so that conversion happens correctly afterwards (otherwise spurious data could be loaded). Then the expression for Q_7 allows P_7 to be overridden by $\overline{E}$ asynchronously. Also, although overkill, the counter bits are all forced to $\texttt{1}$ to ensure conversion occurs correctly. (Really, this should only be needed for C_0). We only need correct behavior if the chip is disabled right after a start

pulse, so the other bits should be okay. However, it is okay to add "$+E$" to the other register inputs if desired. We have also dispensed with D_{Q_D}.

The pin diagrams for the original chip and emulator are shown below. Note that we can now accommodate every output from the original chip.



original 74LS503 **74503 emulator**

The code to implement the state machine in CUPL is shown below.

74503-F750C.pld

```
/*
*   74503 SAR emulator, on an ATF750C
*/
Name            74503-F750C;
Partno          74503;
Revision        01;
Date            5/23/2015;
Designer        Daniel Steck;
Company         University of Oregon;
Location        None;
Assembly        None;
Device          v750c;

/*** inputs ***/
pin 1 = CP;          /* clock pulse (trig on rising edge) */
pin 2 = !S;          /* start low  */
pin 3 = Din;         /* data */
pin 4 = !E;          /* enable low */

/*** outputs ***/
pin 14       = !CC;       /* conversion complete low */
pin [15..22] = [Q0..Q7];  /* 8-bit output */
pin 23       = !Q7copy;   /* inverted copy of Q7 */

/*** internal nodes ***/
node P7;
node [C0..C2];

/*** intermediate counter variables ***/
```

```
C0.D = S # !(!C2 & !C1 & !C0) & !C0 # !E;
C1.D = S # !(C1 & !C0) & C1 # C2 & !C1 & !C0 # !E;
C2.D = S # !(C2 & !C1 & !C0) & C2 # !E;

/*** register inputs ***/
CC.D = !(S # !(!C2 & !C1 & !C0) & !CC);
P7.D = !S & (C2 & C1 & C0 & Din # !(C2 & C1 & C0) & P7) & E;
Q6.D = S # ( C2 &  C1 & !C0 & Din # !( C2 &  C1 & !C0) & Q6) & !( C2 &  C1 &  C0);
Q5.D = S # ( C2 & !C1 &  C0 & Din # !( C2 & !C1 &  C0) & Q5) & !( C2 &  C1 & !C0);
Q4.D = S # ( C2 & !C1 & !C0 & Din # !( C2 & !C1 & !C0) & Q4) & !( C2 & !C1 &  C0);
Q3.D = S # (!C2 &  C1 &  C0 & Din # !(!C2 &  C1 &  C0) & Q3) & !( C2 & !C1 & !C0);
Q2.D = S # (!C2 &  C1 & !C0 & Din # !(!C2 &  C1 & !C0) & Q2) & !(!C2 &  C1 &  C0);
Q1.D = S # (!C2 & !C1 &  C0 & Din # !(!C2 & !C1 &  C0) & Q1) & !(!C2 &  C1 & !C0);
Q0.D = S # (!C2 & !C1 & !C0 & !CC & Din # !(!C2 & !C1 & !C0 & !CC) & Q0) & !(!C2 & !C1 &  C0);

/*** combinatorial outputs ***/
Q7 = !E # P7;
Q7copy = Q7;

/*** handle flip-flop variables set/preset inputs ***/
CC.ar = 'b'0;
C2.ar = 'b'0;
C1.ar = 'b'0;
C0.ar = 'b'0;
P7.ar = 'b'0;
Q6.ar = 'b'0;
Q5.ar = 'b'0;
Q4.ar = 'b'0;
Q3.ar = 'b'0;
Q2.ar = 'b'0;
Q1.ar = 'b'0;
Q0.ar = 'b'0;

CC.sp = 'b'0;
C2.sp = 'b'0;
C1.sp = 'b'0;
C0.sp = 'b'0;
P7.sp = 'b'0;
Q6.sp = 'b'0;
Q5.sp = 'b'0;
Q4.sp = 'b'0;
Q3.sp = 'b'0;
Q2.sp = 'b'0;
Q1.sp = 'b'0;
Q0.sp = 'b'0;

/*** flip-flop-clock multiplexer (use input clock pin) ***/
CC.ckmux = CP;
C2.ckmux = CP;
C1.ckmux = CP;
C0.ckmux = CP;
P7.ckmux = CP;
Q6.ckmux = CP;
Q5.ckmux = CP;
Q4.ckmux = CP;
Q3.ckmux = CP;
Q2.ckmux = CP;
```

```
Q1.ckmux = CP;
Q0.ckmux = CP;
```

Again, this is a relatively straightforward translation of the Boolean-algebraic equations. Note that

- In the first block, we now declare the more powerful chip ("v750c").
- In the next two blocks, we declare $\overline{E}$ and $\overline{Q_7}$ (the latter by defining the "copy" variable `Q7copy`).
- In the next block, we implement the counter variables C_0–C_2 as register variables, as in Eqs. (16.18).
- The next block gives expressions for all the register inputs, as in Eqs. (16.18) (except for the counter register variables we already implemented). The subsequent block implements the combinatorial outputs Q_7 and $\overline{Q_7}$.
- Finally, in the last block, the ATF750C has a "clock multiplexer" control on the flip-flop inputs. The upshot is that we must declare the flip-flop clock inputs to be connected to the clock-input pin `CP`.
- To have $\overline{Q_7}$ as an output instead of D_7, we can change the "`pin 23 = QD;`" declaration to now read "`pin 23 = notQ7;`", and then change the line "`QD.D = Din;`" to "`notQ7 = !Q7;`" (note that this output would then no longer be registered, but "combinatorial").

The compiler then simplifies the logical expressions and figures out how to make the fused connections in the configurable-logic section of the 22V10. A separate programmer is necessary to then "burn" the chip.

16.4.3 Testing the State Machines

The WinCUPL package also allows simulation tests. We define the test values in another file (i.e., define a sequence of input and expected output values). The simulator will simulate the chip and ensure that it passes the test values. These tests values can also be embedded in the code to be sent to the programmer, so the programmer can test the actual chip.

Test files for both emulators are attached below; it's a good exercise to look through and understand these. The header block here matches that of the `.pld` file. The `ORDER` declaration gives a sequence of pins (variables) for consideration. In the `VECTOR` block, we give a bunch of input/expected-output states, in the order of the `ORDER` declaration. The notation for the values is:

- `0` and `1` are the logical input values.
- `L` and `H` are corresponding logical values, but used for expected outputs.
- `c` is equivalent to a `0`, then a `1`, and then a `0` (i.e., a clock pulse).

74502-22V10.si

```
/*
*   74502 SAR emulator, on a 22v10
*/
Name            74502-22V10;
Partno          74502;
Revision        01;
Date            5/23/2015;
Designer        Daniel Steck;
Company         University of Oregon;
Location        None;
Assembly        None;
Device          g22v10;

ORDER: CP, !S, Din, QD, Q7, Q6, Q5, Q4, Q3, Q2, Q1, Q0, !CC;
```

```
VECTORS:
c 00 LLHHHHHHHH
c 10 LLLHHHHHHH
c 10 LLLLHHHHHH
c 10 LLLLLHHHHH
c 10 LLLLLLHHHH
c 10 LLLLLLLHHH
c 10 LLLLLLLLHH
c 10 LLLLLLLLLH
c 10 LLLLLLLLLL

c 01 HLHHHHHHHH
c 11 HHLHHHHHHH
c 11 HHHLHHHHHH
c 11 HHHHLHHHHH
c 11 HHHHHLHHHH
c 11 HHHHHHLHHH
c 11 HHHHHHHLHH
c 11 HHHHHHHHLH
c 11 HHHHHHHHHL
c 11 HHHHHHHHHL

c 01 HLHHHHHHHH
c 10 LLLHHHHHHH
c 11 HLHLHHHHHH
c 10 LLHLLHHHHH
c 11 HLHLHLHHHH
c 10 LLHLHLLHHH
c 11 HLHLHLHLHH
c 10 LLHLHLHLLH
c 11 HLHLHLHLHL
c 10 LLHLHLHLHL

c 01 HLHHHHHHHH
c 11 HHLHHHHHHH
c 10 LHLLHHHHHH
c 11 HHLHLHHHHH
c 10 LHLHLLHHHH
c 11 HHLHLHLHHH
c 10 LHLHLHLLHH
c 11 HHLHLHLHLH
c 10 LHLHLHLHLL
c 11 HHLHLHLHLL
```

74503-F750C.si

```
/*
*   74503 SAR emulator, on an ATF750C
*/
Name            74503-F750C;
Partno          74503;
Revision        01;
Date            5/23/2015;
Designer        Daniel Steck;
Company         University of Oregon;
Location        None;
Assembly        None;
Device          v750c;
```

```
ORDER:
CP, !E, !S, Din, C2, C1, C0, Q7, Q6, Q5, Q4, Q3, Q2, Q1, Q0, !CC;

VECTORS:
c 0 00 HHH LHHHHHHHH
c 0 10 HHL LLHHHHHHH
c 0 10 HLH LLLHHHHHH
c 0 10 HLL LLLLHHHHH
c 0 10 LHH LLLLLHHHH
c 0 10 LHL LLLLLLHHH
c 0 10 LLH LLLLLLLHH
c 0 10 LLL LLLLLLLLH
c 0 10 LLL LLLLLLLLL
c 0 10 LLL LLLLLLLLL

c 0 01 HHH LHHHHHHHH
c 0 11 HHL HLHHHHHHH
c 0 11 HLH HHLHHHHHH
c 0 11 HLL HHHLHHHHH
c 0 11 LHH HHHHLHHHH
c 0 11 LHL HHHHHLHHH
c 0 11 LLH HHHHHHLHH
c 0 11 LLL HHHHHHHLH
c 0 11 LLL HHHHHHHHL
c 0 11 LLL HHHHHHHHL

c 0 01 HHH LHHHHHHHH
c 0 10 HHL LLHHHHHHH
c 0 11 HLH LHLHHHHHH
c 0 10 HLL LHLLHHHHH
c 0 11 LHH LHLHLHHHH
c 0 10 LHL LHLHLLHHH
c 0 11 LLH LHLHLHLHH
c 0 10 LLL LHLHLHLLH
c 0 11 LLL LHLHLHLHL
c 0 10 LLL LHLHLHLHL

c 0 01 HHH LHHHHHHHH
c 0 11 HHL HLHHHHHHH
c 0 10 HLH HLLHHHHHH
c 0 11 HLL HLHLHHHHH
c 0 10 LHH HLHLLHHHH
c 0 11 LHL HLHLHLHHH
c 0 10 LLH HLHLHLLHH
c 0 11 LLL HLHLHLHLH
c 0 10 LLL HLHLHLHLL
c 0 11 LLL HLHLHLHLL

0 0 11 LLL HLHLHLHLL
1 0 00 HHH LHHHHHHHH
1 1 00 HHH HHHHHHHHH
1 1 01 HHH HHHHHHHHH
1 0 10 HHH LHHHHHHHH
0 0 10 HHH LHHHHHHHH
c 0 10 HHL LLHHHHHHH
c 0 10 HLH LLLHHHHHH
```

```
c 0 10 HLL LLLLHHHHH
c 0 10 LHH LLLLLHHHH
c 0 10 LHL LLLLLLHHH
c 0 10 LLH LLLLLLLHH
c 0 10 LLL LLLLLLLLH
c 0 10 LLL LLLLLLLLL
c 0 10 LLL LLLLLLLLL

0 0 00 LLL LLLLLLLLL
1 0 01 HHH LHHHHHHHH
1 1 01 HHH HHHHHHHHH
1 1 00 HHH HHHHHHHHH
1 0 11 HHH LHHHHHHHH
0 0 11 HHH LHHHHHHHH
c 0 11 HHL HLHHHHHHH
c 0 11 HLH HHLHHHHHH
c 0 11 HLL HHHLHHHHH
c 0 11 LHH HHHHLHHHH
c 0 11 LHL HHHHHLHHH
c 0 11 LLH HHHHHHLHH
c 0 11 LLL HHHHHHHLH
c 0 11 LLL HHHHHHHHL
c 0 11 LLL HHHHHHHHL

0 0 01 LLL HHHHHHHHL
1 0 01 HHH LHHHHHHHH
1 1 01 HHH HHHHHHHHH
1 1 00 HHH HHHHHHHHH
1 0 01 HHH LHHHHHHHH
0 0 11 HHH LHHHHHHHH
c 0 10 HHL LLHHHHHHH
c 0 11 HLH LHLHHHHHH
c 0 10 HLL LHLLHHHHH
c 0 11 LHH LHLHLHHHH
c 0 10 LHL LHLHLLHHH
c 0 11 LLH LHLHLHLHH
c 0 10 LLL LHLHLHLLH
c 0 11 LLL LHLHLHLHL
c 0 10 LLL LHLHLHLHL

0 0 00 LLL LHLHLHLHL
1 0 01 HHH LHHHHHHHH
1 1 01 HHH HHHHHHHHH
1 1 00 HHH HHHHHHHHH
1 0 01 HHH LHHHHHHHH
0 0 11 HHH LHHHHHHHH
c 0 11 HHL HLHHHHHHH
c 0 10 HLH HLLHHHHHH
c 0 11 HLL HLHLHHHHH
c 0 10 LHH HLHLLHHHH
c 0 11 LHL HLHLHLHHH
c 0 10 LLH HLHLHLLHH
c 0 11 LLL HLHLHLHLH
c 0 10 LLL HLHLHLHLL
c 0 11 LLL HLHLHLHLL
```

16.5 Circuit Practice

16.5.1 Computer-Interface DAC Controller

For circuit practice, see the DAC controller board design by Todd Meyrath and Florian Schreck.[4] Trace through the circuit and note the following.

- The DAC7744 chips have 4 analog outputs, for 8 total output channels per board.
- An 8-bit address bus selects which DAC and output to use. The two LSBs (bits 0 and 1) select which output on a particular chip, bit 2 selects which of the two DACs on the board to activate, and the other bits select which (of possibly many) boards to address. Trace through the logic leading up to and including the `NAND` gates to verify that it works as advertised.
- This allows only one 16-bit data bus to feed all the outputs. The desired output is selected, the desired data is presented to the data bus, and then a strobe signal causes the addressed DAC to latch the desired output value.
- Note that the strobe pulse, which just amounts to matching the proper address, must be delayed behind the data and address signals so the inputs are settled before "load DAC" is triggered. The pulse is delayed by a buffered, RC circuit.
- Note that the `NOT` gates have Schmitt-trigger inputs. What part of the circuit justifies having Schmitt-input `NOT`s?

16.5.2 3-Bit ADC

Suppose you have a 3-bit DAC, with voltage levels 0 V, 0.1 V, 0.2 V, ..., in a successive-approximation ADC. If $V_{\text{in}} = 0.35\,\text{V}$,

- make a plot of the DAC output vs. time
- what is the final, converted digital value?

Solution. For the converted digital value: `100`. The comparison voltages will be: 0.3 V, 0.5 V, 0.4 V.

[4] `http://strontiumbec.com/Control/DAC.pdf`, in particular the schematic on p. 13.

16.6 Exercises

Problem 16.1

(a) Derive an expression for the dynamic range (the largest vs. the smallest nonzero-amplitude signal) of an N-bit sampled signal. Recall that when you compare two amplitudes A and A_0 in dB, the expression is

$$(\text{ratio in dB}) = 20 \log_{10} \left(\frac{A}{A_0} \right) \tag{16.19}$$

(b) What is the dynamic range for CD audio (16 bits) in dB? Bluray audio (24 bits)? (For comparison, the dynamic range of human hearing is usually quoted as 120 dB.)

Problem 16.2

Use a counter to design a simple DAC based on pulse-width modulation as follows: Your circuit should take an 8-bit digital input, representing an 8-bit unsigned integer, and then control the brightness of an LED to be proportional to this integer. Assume the clock signal to be given and to be as fast as you need it to be. You may use whatever support logic you like, but you may find it useful to use a flip-flop, and look into the binary magnitude comparator (read up on the 74688). Make sure to properly limit the LED current.

Be *specific* about any ICs you use (i.e., give the model number, like 74688, and if it matters, specify which logic family, e.g., 74HCT688). You should show all important connections, but don't bother with universal stuff like power supplies, grounds, chip enables, etc. Also, you don't need to explicitly show the clock source, just indicate the existence of the clock signal and show any connections where it enters your circuit.

Problem 16.3

(a) Use a counter (specifically, an 8-bit up counter), a clock source (astable multivibrator), a comparator, and a DAC to design a simple ADC. That is, the counter should count upwards starting from zero, and freeze at the appropriate conversion value (i.e., the frozen counter is the output). Use whatever support logic you like, and don't worry about latching the output. You should include a start/reset input that resets the counter and allows the next conversion to start.

As in Problem 2, be *specific* about any ICs you use; show all important connections, but don't bother with universal stuff; and you don't need to explicitly show the clock source, just indicate the existence of the clock signal and show any connections where it enters your circuit.

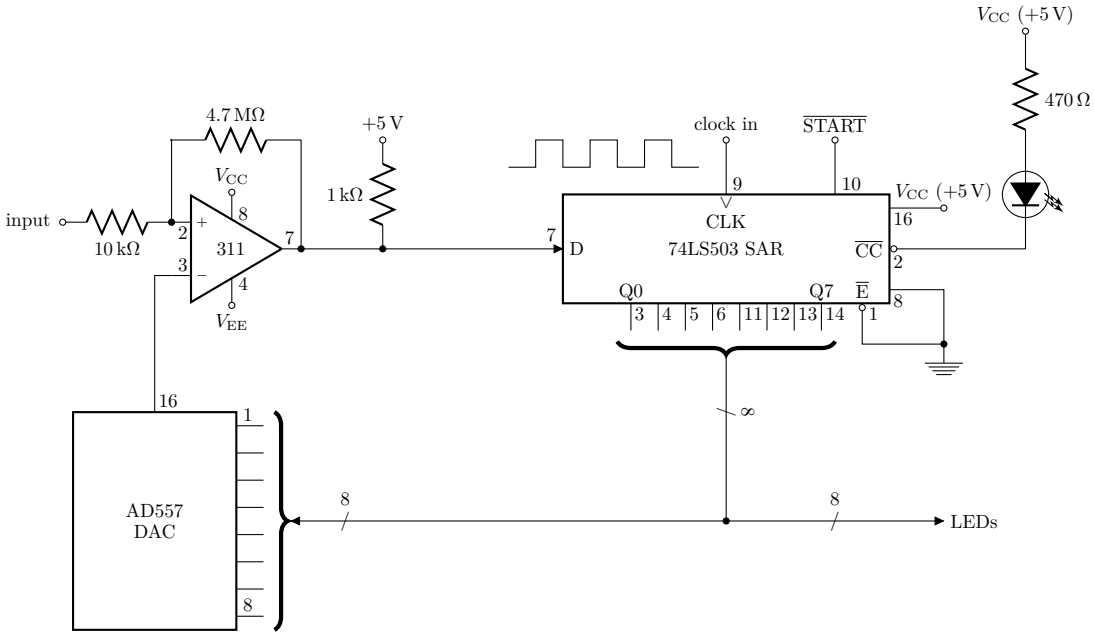
(b) Why is this ADC slower than a successive-approximation ADC? (And by how much is this ADC slower on average?)

(c) What kind of "bias" does this ADC have in terms of converting a noisy input signal?

Problem 16.4

(a) In the successive-approximation ADC examples from Section 16.2.2, the convention is that the SAR starts with the "midpoint" word `01111111` for an 8-bit ADC (as in the 74LS502 and 74LS503 SAR's in Section 16.4). Another possible (and reasonable) convention is to start off with the the *alternate* midpoint word `10000000`. Briefly describe the difference in the *end result* of the two schemes for an arbitrary input.

(b) Consider the ADC circuit shown below, where the DAC conversion levels are 0, 0.01, 0.02, ..., 2.55 V. Suppose the input analog voltage of 0.45 V, and as mentioned the SAR starts with `01111111` on the first clock cycle (the "start" cycle). What is the SAR output after 3 more clock cycles?

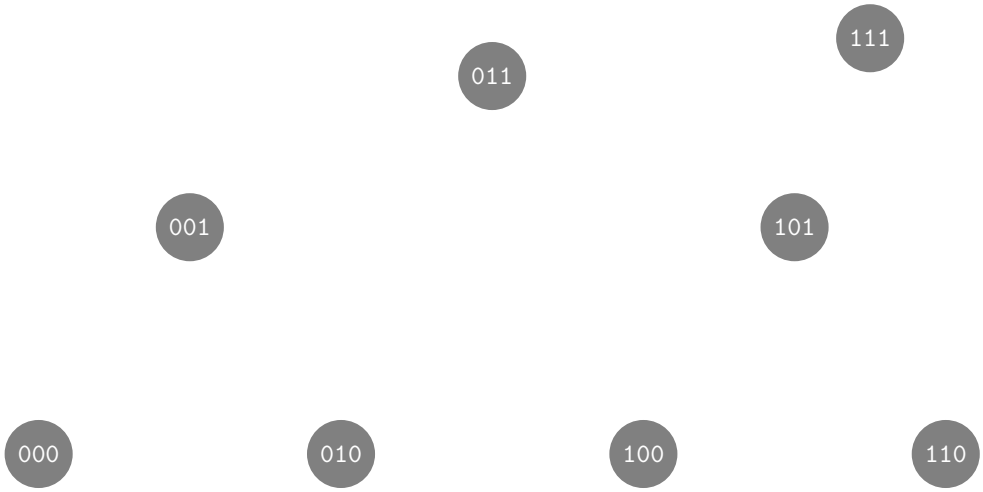


Problem 16.5

In this problem, you should design a 2-bit successive-approximation register (SAR), with data (D) and start-LOW ($\overline{S}$) inputs, Q_1 (MSB) and Q_0 (LSB) outputs, and conversion-complete-LOW ($\overline{CC}$) output. On start, the outputs should initialize to $Q_1Q_0\overline{CC} = 011$.

(a) Draw a state diagram for the SAR, enumerating all possible output states $Q_1Q_0\overline{CC}$ as nodes. Specify **all** possible transitions, and label transition arrows with the appropriate input states wherever multiple transitions are possible. Also, make sure to handle all possible states and eliminate the possibility that the state machine will get "stuck." (Remember you can use "X" for "doesn't matter" for logic states in diagrams and truth tables.)

A suggested template for your state diagram is shown below. (That is, these are the states, you should fill in the transitions.)



(b) Write down a truth table for register inputs D_1, D_0, and $D_{\overline{CC}}$ in terms of the other variables to implement this SAR in sequential logic.

Again, a suggested template for your solution is shown below.

$\overline{S}$	D	Q_1	Q_0	$\overline{\texttt{CC}}$	D_1	D_0	$D_{\overline{\texttt{CC}}}$
0	X	X	X	X			
X	X	1	1	1			
1	0	0	1	1			
1	1	0	1	1			
1	0	0	0	1			
1	1	0	0	1			
1	0	1	0	1			
1	1	1	0	1			
1	X	X	X	0			

(c) Finish the design: write down logical expressions for D_1, D_0, and $D_{\overline{\texttt{CC}}}$.

Chapter 17

Phase-Locked Loops

Simply put, a **phase-locked loop (PLL)** is a feedback-loop circuit that compares two oscillating signals. It attempts to adjust the frequency of the second one so that it exactly matches the first in terms of phase (and thus also in terms of frequency).

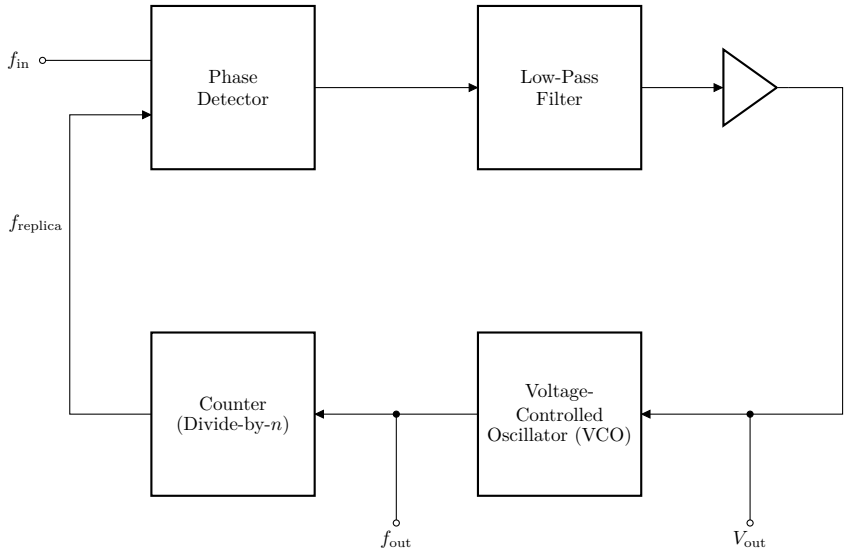
Strictly speaking, a phase-locked loop can be implemented in an analog circuit, where, for example, the circuit makes one sine wave copy another one. However, it is common to implement phase-locked loops using digital gates, so we are covering these as digital circuits.

It may also sound a bit weird to use a feedback loop to make a copy of a signal, when you could just directly make a copy of a signal, e.g., with a buffer amplifier or gate. However, the magic comes in taking advantage of the feedback loop. As our first main example, recall that using a counter, it is relatively straightforward to *divide* the frequency of a square-wave clock signal. But how do we *multiply* the frequency of a signal? The answer: a phase-locked loop.

17.1 Frequency Multiplier

The idea behind a frequency multiplier is to start with the original clock signal. Suppose we want to multiply the frequency by N. Then generate a new signal, *divide it by* N, and compare the *divided* signal to the original (i.e., phase synchronized). Adjust the frequency of the new signal until the divided version matches the original, and *voila*, you have a new signal with a frequency N times the original, with matching phases of the two signals.

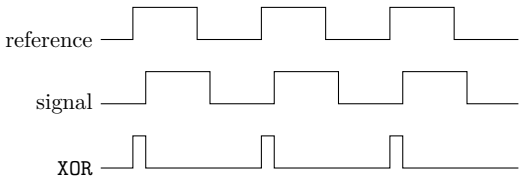
The block diagram of a circuit that accomplishes this is shown below.



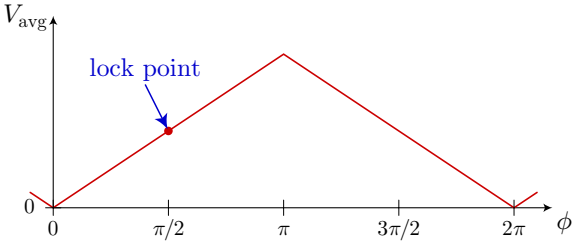
Let's look at each of the components here.

1. The **phase detector** compares two oscillating signals, and the output gives some measure of the relative phase. A general requirement is that the "in sync" state should give a "zero output"—this need not actually be zero (i.e., it could be offset to some other voltage), but the point is that the signal should go up if the phase is perturbed one way, and down if the phase is perturbed the other way. There are two basic classes of digital phase detectors in PLLs.

   - **Type I phase detector**. This is simply an `XOR` gate, and it can be driven by digital signals, or also by analog signals, provided they have been converted to digital via a comparator or Schmitt trigger. The output of the `XOR` gate is illustrated below.

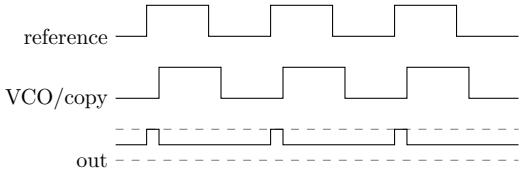
     The output is a sequence of pulses; the output is `HIGH` whenever the two input signals mismatch. The `XOR` output is zero if the reference input matches the signal input, and the duty cycle of the output increases to 100% if the signals have a π phase difference. Only the *average* signal will matter, because the output is fed through a low-pass filter. So the average output is proportional to the phase difference, as shown below.

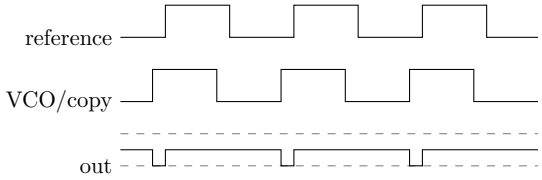
     Then we should choose a $\pi/2$ phase shift as the lock point (i.e., the PLL will force the signal to be a 90°-phase-shifted copy of the reference). Remember this is because we need the output signal to vary both up *and* down if the phase moves away from the lock point.

   - **Type II phase detector**. This phase detector is sensitive to digital *edges*, so it is really suited to digital signals, although in principle if analog signals are converted to digital, this would amount to

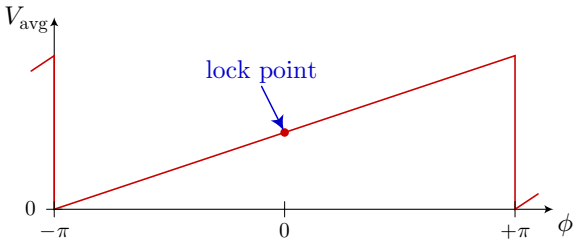
the detector responding to the zero-crossings of the analog signals. To see how it works, consider the first timing diagram below.



The "normal" state of the phase detector is a *middle* voltage midway between `LOW` and `HIGH`. If the detector finds a rising edge from the reference signal first, then the output changes `HIGH`. When the edge from the other signal ("VCO/copy" signal, which we will explain below), then the output changes back to `MID`. If the relative phase has the opposite sign, then we have the situation shown below.



Now the VCO/copy signal presents its rising edge first, so the output goes from `MID` to `LOW`. It goes back to `MID` when the reference edge arrives. Once this signal is time-averaged, the result is shown below.



Now the lock point is at zero phase, and because the output can move `LOW` or `HIGH` relative to the normal `MID` state, the output can vary in either direction. The advantage is that the lock point is in perfect sync: at the lock point, the error signal is identically zero, ever before the time average. Contrast this to the type-I case, where the output was a 50%-duty-cycle square wave. Some of this will leak through the time average, and end up frequency-modulating the output signal. Note that the operation of this circuit is independent of the duty cycles of the two signals, unlike the type-I case where we assumed 50% duty cycles. (Otherwise the locked phase may differ from $\pi/2$, and it may be necessary to choose a different lock voltage.)

Note that for sine-wave analog signals, the phase detector can be as simple as a multiplying amplifier (for rf frequencies, you would use an **rf mixer**, which has just this function).

2. The **low-pass filter** (Section 2.3.5) "keeps" low frequencies, and "removes" high frequencies. It thus acts to time-average the phase-detector signal. It also limits the speed with which the PLL can respond to frequency changes in the reference signal. This may be a disadvantage if you want to perfectly track the frequency. However, this allows the PLL to act as a frequency "flywheel," so it ignores some of the noise in the incoming signal to "clean it up." It also induces a phase shift, which we will return to below.

3. The **voltage-controlled oscillator (VCO)** is an oscillator (clock), where the output frequency f depends on an input control voltage. The frequency may depend nonlinearly on the control voltage, but it should be at least monotonic.

4. The **counter** is here as we described for frequency-multiplier applications. This should be omitted in other applications.

17.1.1 Feedback Loop

In the feedback loop here, the time-averaged output of the phase detector is fed into the VCO control input. Recall that in PID control (Chapter 8), it is necessary to integrate the error signal in order to get zero steady-state error. Here, this is automatic, since frequency ω and phase ϕ are related by

$$\omega = \frac{d\phi}{dt}. \tag{17.1}$$

Since we detect phase and are feeding back to a frequency control, we are controlling $\omega = \int \phi \, dt$, and thus we effectively have the integral of the error signal.

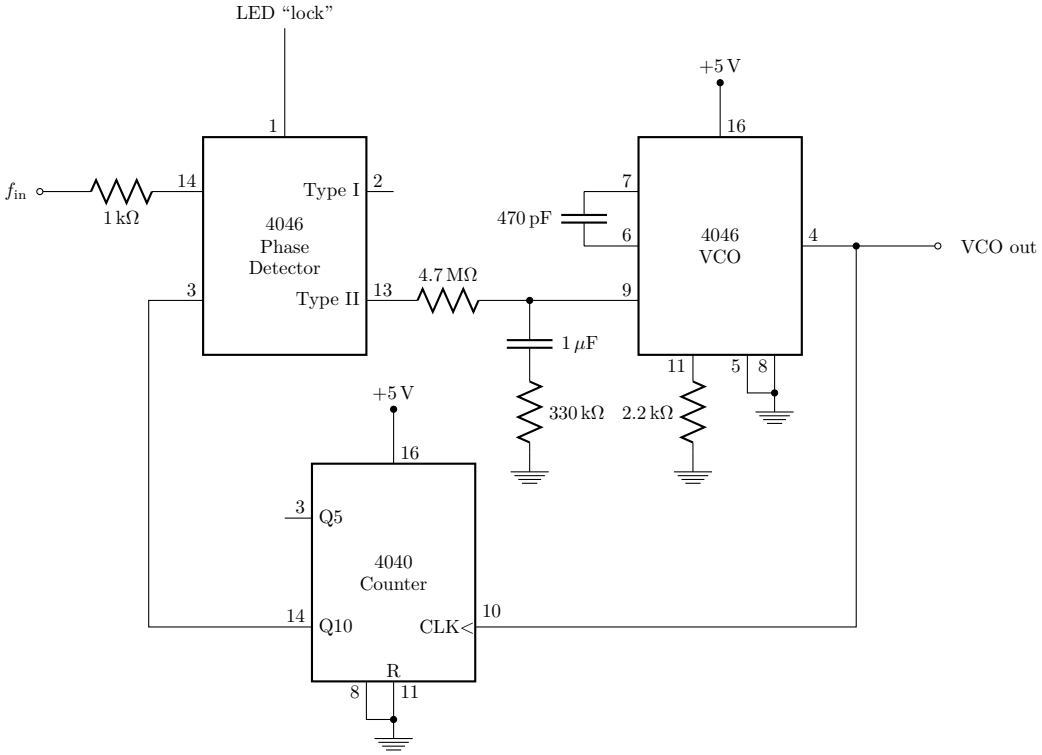
If the phase is *below* the lock point (i.e., phase lag), then the output is positive (relative to the lock point), and so the frequency increases. If the phase is *above* the lock point (i.e., phase lag), then the output is negative (relative to the lock point), and so the frequency decreases. Then there are different options for loops.

1. In a **first-order loop**, there is no low-pass filter, so there is just the 90° phase shift associated with the phase-frequency integration. In this case, we don't have the time averaging as in the analysis above, but the idea is the same, because the integral VCO response makes the loop behave in essentially the same way.

2. In a **second-order loop**, there *is* a low-pass filter as in the diagram, so you need to be careful about any extra phase shifts to guard against instabilities. Again, the low-pass filter limits the rate of change (i.e., the bandwidth) of the control signal.

17.2 Example PLL

Below is a more detailed example of a PLL circuit, based on the 4046 PLL IC. This IC includes both type I and type II phase detectors, as well as a VCO. The circuit below is designed to multiply an input 60-Hz signal by 2^5 or 2^{10}, depending on whether the Q_5 or Q_{10} output of the 4040 counter is used. Note the passive low-pass filter and the use of the type II detector, so this is a second-order feedback loop.



17.3 Other Applications

17.3.1 FM Demodulation

Another important application of PLL circuits is frequency-modulation (FM) demodulation (i.e., the demodulator in an FM radio). In this case, the receiver receives an FM signal, and a PLL attempts to reproduce this signal. The FM signal was generated by changing the frequency according to some signal to be transmitted (e.g., audio). Then the control voltage to the VCO is a copy of the original signal, provided the VCO control voltage is related to frequency in the same way as in the original FM process. Typically, this just means that the VCO frequency should be linear in the control voltage.

17.3.2 Direct Digital Synthesis

Another application is in **direct digital synthesis (DDS)**. The idea here is to take a precision clock input (e.g., from an atomic clock or oven-stabilized crystal oscillator), multiply it to a high frequency, and then use a counter to divide it to some other frequency. This can produce frequencies with high resolution over a wide range if the frequency-multiplication factor is large. The divider is digitally programmable so the final frequency is dynamically programmable. Then the digital counter output drives an analog "look-up table" of voltages to get a high-quality (low-distortion), timing-accurate sine wave. An example is the AD985L DDS IC, which can take a 10-MHz clock in, and produce a sine-wave output in the range of 0–135 MHz.

17.4 Dynamical Model

To understand the behavior of a phase-locked loop in more depth, here we will develop a simple dynamical model. Consider a reference signal $V_{\rm ref}$, given by

$$V_{\rm ref}(t) = V_{\rm r0}\cos\phi_{\rm ref}(t), \tag{17.2}$$

where the reference frequency is

$$\omega_{\rm ref} = \dot{\phi}_{\rm ref}. \tag{17.3}$$

We will take this frequency to be constant , so that $\phi_{\rm ref} = \omega_{\rm ref}t$. The signal that we want to phase-lock to the reference is similarly

$$V_{\rm sig}(t) = V_{\rm s0}\sin\phi_{\rm sig}(t), \tag{17.4}$$

where the signal frequency is

$$\omega_{\rm sig}(t) = \dot{\phi}_{\rm sig}(t), \tag{17.5}$$

or inverting this relation,

$$\phi_{\rm sig}(t) = \int_0^t \omega_{\rm sig}(t')\,dt'. \tag{17.6}$$

Note that we have already built in a relative phase of $\pi/2$ between reference and signal, anticipating that the two signals will prefer to lock with this phase difference. Thus, the locking condition is that $\phi_{\rm sig}(t) = \phi_{\rm ref}(t)$, modulo 2π.

The simplest phase detector for the analog signals here is a multiplier for two analog signals, which is called a **mixer** for radio-frequency (rf) signals. This is the analog equivalent of the Type-I phase detector (the `XOR` gate). Think of the `XOR` gate operating on logic state of 1 and -1. The `XOR` (or product) is 1 if two input signals are the same, and the `XOR` (or product) is -1 if the two signals are opposite. We can write the output of the mixer as

$$V_{\rm mix}(t) = \frac{V_{\rm ref}(t)\,V_{\rm sig}(t)}{V_0} = \frac{V_{\rm r0}\,V_{\rm s0}}{V_0}\cos\phi_{\rm ref}\sin\phi_{\rm sig}. \tag{17.7}$$

Then defining $V_{\rm m0} := V_{\rm r0}V_{\rm s0}/2V_0$ and using the identity $\sin\alpha\cos\beta = (1/2)[\sin(\alpha-\beta)+\sin(\alpha+\beta)]$, the mixer signal becomes

$$V_{\rm mix}(t) = V_{\rm m0}\Big[\sin(\phi_{\rm sig}-\phi_{\rm ref}) + \sin(\phi_{\rm sig}+\phi_{\rm ref})\Big]. \tag{17.8}$$

In phase-locked-loop operation, only the first term will be important here. At or near the locking condition, $\omega_{\rm sig} \approx \omega_{\rm ref}$, so the first term varies slowly (close to dc), while the second term has a frequency of $\omega_{\rm sig} + \omega_{\rm ref} \approx 2\omega_{\rm ref}$, which is much faster. The effect of this fast oscillation will tend to be averaged away to zero, especially as we will be feeding this mixer signal through a low-pass filter, which will greatly suppress the second term. Thus, we will write

$$V_{\rm mix}(t) = V_{\rm m0} \sin(\phi_{\rm sig} - \phi_{\rm ref}), \tag{17.9}$$

as far as the operation of the feedback loop is concerned.

To complete the loop, we must connect the output of the mixer to the input of the VCO, via a low-pass filter (to make a second-order loop). Then the VCO outputs a signal at frequency

$$\omega_{\rm sig}(t) = \omega_{\rm s0} - \frac{g_{\rm I}}{V_{\rm m0}} \int_0^t V_{\rm mix}(t')\, dt', \tag{17.10}$$

where $\omega_{\rm s0}$ is the "natural" frequency of the VCO (i.e., the frequency with zero input voltage), $g_{\rm I}$ is an "integral" gain factor, and we are modeling the low-pass filter via an integral (recalling from Section 2.2.1 that a low-pass filter acts as an integrator provided the output signal is small compared to the input). Note that we have included a minus sign here to provide negative feedback. However, this is optional, but without it the relative phase at the lock point will differ by π from the analysis here.

17.4.1 Equation of Motion

Now we can write down a dynamical equation for the phase-locked loop. Consider the phase difference

$$\Delta\phi(t) := \phi_{\rm sig}(t) - \phi_{\rm ref}. \tag{17.11}$$

The first derivative gives the frequency difference

$$\Delta\dot{\phi}(t) = \omega_{\rm sig}(t) - \omega_{\rm ref}. \tag{17.12}$$

Putting in Eq. (17.10) for $\omega_{\rm sig}$,

$$\Delta\dot{\phi}(t) = \omega_{\rm s0} - \omega_{\rm ref} - \frac{g_{\rm I}}{V_{\rm m0}} \int_0^t V_{\rm mix}(t')\, dt'. \tag{17.13}$$

Differentiating this equation

$$\Delta\ddot{\phi}(t) = -\frac{g_{\rm I}}{V_{\rm m0}} V_{\rm mix}(t), \tag{17.14}$$

and then using Eq. (17.9),

$$\Delta\ddot{\phi}(t) = -g_{\rm I} \sin(\Delta\phi). \tag{17.15}$$

This equation has the form of a mechanical pendulum, $\ddot{\theta} = -(g/\ell)\sin\theta$. This means that $\Delta\phi = 0$ (modulo 2π) is a steady state, meaning that once locked, the circuit can stay locked. However, there is no means to *become* locked: if the relative phase is displaced from the lock point, it will oscillate back and forth about it without settling. Thus, we need to introduce some *damping*.

17.4.2 Damping

To introduce damping, we will introduce a "proportional" term in the feedback in Eq. (17.13):

$$\Delta\dot{\phi}(t) = \omega_{\rm s0} - \omega_{\rm ref} - \frac{g_{\rm I}}{V_{\rm m0}} \int_0^t V_{\rm mix}(t')\, dt' - \frac{g_{\rm P}}{V_{\rm m0}} V_{\rm mix}(t). \tag{17.16}$$

Here $g_{\rm P}$ is the "proportional" gain. This model corresponds to the two-resistor, one-capacitor filter in the example PLL circuit on p. 326. At low frequencies, the filter acts as an ordinary low-pass filter (integrator), while at high frequencies the capacitor acts as a short, so the two resistors form a divider that determine $g_{\rm P}$.

Now putting this feedback into Eq. (17.12), the result is

$$\Delta\ddot{\phi}(t) = -g_{\text{I}} \sin(\Delta\phi) - g_{\text{P}} \Delta\dot{\phi} \cos(\Delta\phi), \tag{17.17}$$

in place of Eq. (17.15). The new equation has the form of a damping term. We should compare this equation of motion to that of the damped pendulum,

$$\ddot{\theta} = -\frac{g}{\ell} \sin\theta - \frac{\gamma}{\ell}\dot{\theta}, \tag{17.18}$$

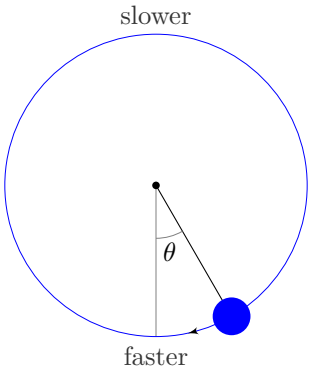
where γ is the damping rate (i.e., the last term here applies a torque that opposes the angular velocity, slowing the pendulum). The PLL equation (17.17) has a similar damping term, but modulated by the cosine of the relative phase. It isn't completely clear that this helps: the sign of the damping changes depending on the phase, and a negative damping is no damping at all (it tries to *speed* up the rate of phase change).

If we look close to lock, then this equation works out. That is, suppose $\Delta\phi$ is small. Then $\sin(\Delta\phi) \approx \Delta\phi$, and $\cos(\Delta\phi) \approx 1$. Thus, Eq. (17.17) becomes

$$\Delta\ddot{\phi}(t) \approx -g_{\text{I}} \Delta\phi - g_{\text{P}} \Delta\dot{\phi}, \tag{17.19}$$

which is the equation for a damped harmonic oscillator. Thus, $\Delta\phi$ will settle to zero (or a multiple of 2π), possible exhibiting damped oscillations about the lock point along the way.

If the circuit is far from lock, then $\Delta\phi$ varies rapidly, and it isn't clear that the damping helps, because $\cos(\Delta\phi)$ seems like it should average to zero. The key to understanding how the damping works is to consider a mechanical pendulum, as shown below, rotating at high angular velocity.



Because the kinetic energy is lower when the pendulum is going "over the top," the angular velocity is lower at the top than at the bottom. In the phase-locked loop, this means $|\Delta\dot{\phi}|$ is larger when $\Delta\phi$ is near $0, \pm 2\pi, \pm 4\pi, \ldots$, while it is smaller when $\Delta\phi$ is near $\pm\pi, \pm 3\pi, \ldots$ We can model this by writing (assuming $\Delta\dot{\phi} > 0$)

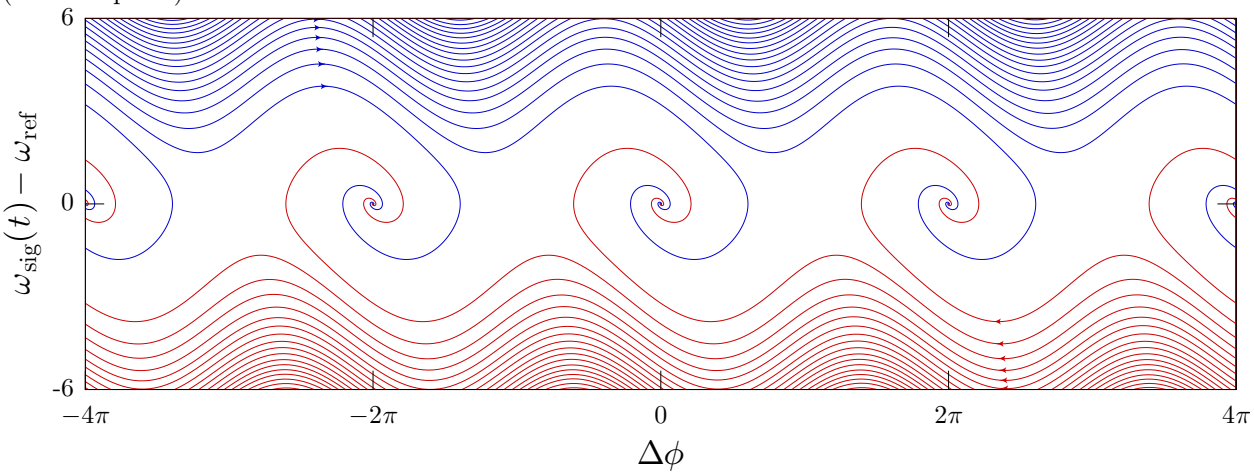
$$\Delta\dot{\phi} = (\omega_{\text{sig}} - \omega_{\text{ref}}) + \delta\omega \cos(\Delta\phi), \tag{17.20}$$

where we assume $(\omega_{\text{sig}} - \omega_{\text{ref}})$ to be slowly varying (i.e., constant on the time scale that $\Delta\phi$ rotates through 2π), and $\delta\omega$ is small. Then the damping term from Eq. (17.17) becomes

$$-g_{\text{P}} \Delta\dot{\phi} \cos(\Delta\phi) = -g_{\text{P}} (\omega_{\text{sig}} - \omega_{\text{ref}}) \cos(\Delta\phi) - g_{\text{P}} \delta\omega \cos^2(\Delta\phi). \tag{17.21}$$

Since $\Delta\phi$ changes rapidly, the first term on the right-hand side will average to zero (because the average value of $\cos x$ is zero), while the second term on the right will average to $-g_{\text{P}}\delta\omega/2$ (because the average value of $\cos^2 x$ is $1/2$). This leads to a net damping that opposes $\Delta\dot{\phi}$. Recalling that $\Delta\dot{\phi}$ is the relative frequency $\omega_{\text{sig}}(t) - \omega_{\text{ref}}$, this means that the damping pushes $\omega_{\text{sig}}(t)$ towards ω_{ref}, until it is close enough that the phase-locked loop can "capture" the signal, and then the phase stabilizes. This damping effect becomes smaller the further ω_{sig} is from ω_{ref}, so if the signal frequency is initially highly mismatched, it may take a while for the loop to attain lock.

The process is illustrated in the plot below, which shows the relative frequency $\Delta\dot{\phi}$ plotted against the relative phase itself, from numerical solutions of Eq. (17.17), with $g_{\text{P}} = 1$ and $g_{\text{I}} = 2$. The trajectories starting far away from the correct frequency are pushed towards the correct frequency, with "bumps" in the frequency along the way, and the capturing process is evident, where the trajectories settle down to a point (the lock point).



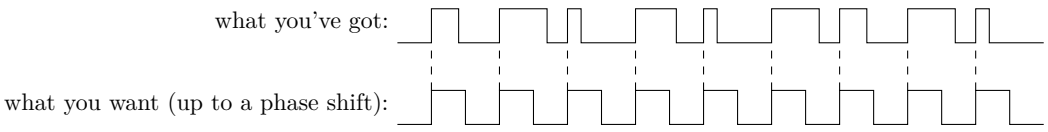
17.5 Exercises

Problem 17.1

Recall that a phase-locked loop based on a Type I detector is sensitive to the duty cycle of the input signal.

(a) **Briefly**, why?

(b) Suppose you have a signal consisting of a train of (digital) pulses. The *rising* edges occur at regularly, at a well-defined frequency. The *falling* edges, however, occur at irregular times. Show (**draw a schematic and describe your reasoning**) how to use two flip-flops (pick your favorite type) as divide-by-2 counters with a phase-locked loop to create a "cleaned" version of the same signal (i.e., square wave, 50% duty cycle, same frequency as the rising edges, "ignores" the falling edges). Note that the actual output may be phase-shifted compared to what is shown here.



You can use the following schematic symbols for the PLL components in your solution (V_{cont} = control voltage; f_{out} = oscillator output signal):

You can assume the circuit will work without a low-pass filter.

Index