Experiment 3

Nonlinear circuits: diodes and analog multipliers

Experiment 3

Nonlinear circuits: diodes and analog multipliers

So far the analog circuits we have considered have all been linear, so that the output has been given by a sum of terms, each term strictly proportional to only one source value (see the section in Experiment 1: ***Linear circuits and superposition*** on page 1-20, and, in particular, equation 1.7 on 1-21). It's now time to extend our design toolbox to include *nonlinear* elements and networks, ones for which the *principle of linear superposition* no longer holds.

The simplest nonlinear circuit component is the *semiconductor diode*, which we consider first. This two-terminal element behaves in a most asymmetric manner: its resistance is very low for currents of more than a few milliamps flowing in one direction through the device, but it has an enormously high resistance to current flow in the opposite direction. This element is very useful for constructing absolute value, peak detection, overvoltage protection, and more general nonlinear resistance circuits. The diode's current-voltage relationship is actually exponential, so it is also useful for building exponential and logarithmic response amplifiers. Its characteristics are strongly temperature-dependent, so a diode also makes an excellent, accurate temperature sensor. Of course, some types of diodes also can emit and detects light (LEDs, laser diodes, and photodiodes), so the variety of applications of the seemingly simple semiconductor diode is nearly endless.

The other nonlinear element we'll consider in this experiment is the much more sophisticated *analog multiplier* integrated circuit, whose output is proportional to the *product* of two input voltages (making its transfer function a so-called *bilinear form* of its two inputs). This flexible device may be used in circuits which not only multiply but also divide, raise to powers, and take roots. You may use it to build voltage-controlled filters and variable-gain amplifiers, modulators and demodulators, phase detectors, automatic gain control and signal compression circuits, and oscillators — as well as its obvious applications in general *analog computing* circuits.

APPLICATIONS OF THE SEMICONDUCTOR DIODE

Diode rectifier circuits

A semiconductor *diode* is a two-terminal element which acts as a "one-way valve" for electrical current (i.e., it is a *rectifier*). The most common type of diode is made from a silicon crystal divided into two layers with different impurity atoms mixed into the silicon. The resulting structure creates a *PN junction* at the interface between the two layers which gives the diode its rectification property. A very simplistic, qualitative description of this rather complicated phenomenon of solid-state physics is provided in the section **THE PN JUNCTION DIODE** on page 3-26.

The schematic symbol of a diode is shown at right, along with a photo of a typical silicon signal diode. When the diode is *forward-biased* its resistance becomes very small, and current will flow through it in the direction shown (note that the schematic symbol includes an arrow (triangle) which points in the direction of the current flow). Forward-biasing is accomplished by applying a voltage so that the diode's *anode* is at a more positive (+) voltage than its *cathode*, as shown in the figure. As you can see in the figure, *the cathode is denoted by a line in the schematic symbol and is usually marked by a line or stripe on the physical diode's body*.

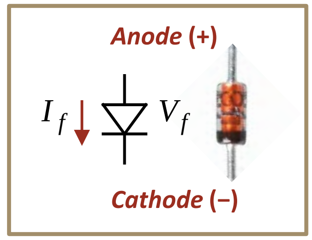


Figure 3-1: A typical silicon signal diode (type 1N4148) and its schematic symbol. The diode's glass case is actually only about 3 mm long.

When a diode is forward-biased and conducts a current of at least a few milliamps, the voltage drop across its two terminals (V_f in Figure 3-1) remains very nearly constant even for an increase in the forward current (I_f) of an order of magnitude or more. When a diode is *reverse-biased* (anode more negative than the cathode), on the other hand, only a very small *leakage current* flows through it. This leakage current is quite insensitive to the reverse voltage applied to the diode, at least until some critical reverse voltage is reached which causes the diode to suddenly *break down* (and, usually, catastrophically fail). Given this basic behavior of the diode as a rectifier, we can construct the following first approximation of its characteristics (good enough to use for many applications):

LOWEST-ORDER DIODE CHARACTERIZATION

A diode's basic behavior is characterized by the following two parameters:

V_f ***forward voltage drop***: the approximately constant voltage drop across the diode when it conducts current in its forward-biased direction.

I_R ***reverse leakage current***: the small current which flows through the diode when the applied voltage is less than V_f or whenever the diode is reverse-biased.

The forward voltage drop, V_f, of a *PN* junction diode is determined by the semiconductor material from which it is constructed; in the case of silicon (Si) this voltage is approximately 0.5–0.7V, whereas for germanium (Ge) V_f is 0.3–0.4V, and for LEDs V_f is >1.7V. The reverse leakage current, I_R, may vary by a couple of orders of magnitude depending on the type of diode and its temperature, but is generally a few tenths of a microamp or less (and may be much less). For a *photodiode* this reverse leakage current increases linearly with the intensity of any light shining on it. A *perfect diode* would be one with $V_f \equiv 0$ and $I_R \equiv 0$.

> A *forward-biased* diode is said to be *on* or *conducting*. A *reverse-biased* diode is said to be *off*.

Basic diode rectifier circuits

Two common diode rectifier circuits are shown in Figure 3-2; we will analyze each using our simple diode model. The input source to each circuit provides an AC signal $v_{in}(t)$; the

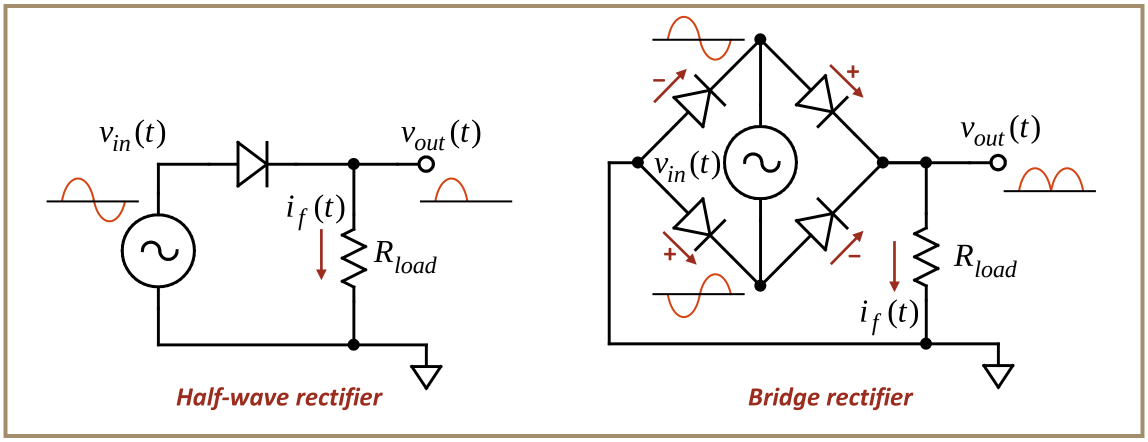


Figure 3-2: Diode rectifier circuits. The AC driving source is converted to a *rectified* output (always of the same polarity); with the chosen orientation of the diodes $v_{out}(t) \geq 0$. As shown by the waveform plots, only the positive half of an AC input cycle gets to the output of the *half-wave rectifier*, but the *bridge rectifier* routes both halves of an AC cycle to the output. A major drawback of the bridge, however, is that the input source must *float* (be isolated from ground potential), whereas the half-wave rectifier allows the source and the load (R_{load}) to share a ground connection.

circuits *rectify* the input to produce an output $v_{out}(t)$ across the load R_{load} which is always of the same polarity (in this case, $v_{out}(t) \geq 0$). We will analyze the action of the *half-wave rectifier* circuit first.

The half-wave rectifier (left-hand circuit in Figure 3-2) is simple to analyze, particularly if we make the assumption that the diode is perfect (both V_f and $I_R \equiv 0$). Whenever $v_{in}(t) > 0$, the diode's anode becomes positive, and it conducts current. The current from the source flows through R_{load}, and, since we assume $V_f = 0$, $v_{out} = v_{in}$. When $v_{in}(t) < 0$, the polarity across the diode reverses, and it turns off; thus $i_f = 0$, and therefore $v_{out} = R_{load}\, i_f = 0$. So the output waveform will appear as in the figure: only the positive half of an input cycle reaches the output.

The *bridge rectifier* (right-hand circuit in Figure 3-2) is more complicated but also more efficient. Assume the diodes are perfect, and consider first the case where the top terminal of the input source is positive with respect to its bottom terminal ($v_{in} > 0$). In this case the two diodes with the "**+**" current arrows can conduct (the other two will be off), and current from the top source terminal can flow through the upper-right diode, the load, and back through the lower-left diode to reach the bottom source terminal. Thus (since we assume the diodes' forward voltage drop vanishes), $v_{out} = v_{in}$. When the polarity of the source is reversed, the two diodes marked with "**–**" current arrows turn on instead, but the current from the source must still flow through R_{load} in the same direction as before, and now $v_{out} = -v_{in}$. Both halves of the source AC cycle are routed to the load, and the output waveform is the absolute value of the input waveform.

> Note the very specific orientation of the diodes in the bridge circuit in Figure 3-2; the roles of input and output may not be exchanged! In particular, the orientations of the pair of diodes at either of the two output terminals are the same: both cathodes attached to the positive output terminal and both anodes attached to the negative output terminal. This is an easy thing to get wrong when you attempt construct a diode bridge.

If the diodes aren't perfect in the two rectifier circuits, then a forward voltage drop V_f will be lost across each conducting diode (V_f is also commonly called a *diode drop*). Thus in the half-wave rectifier when the diode conducts, $v_{out} = v_{in} - V_f$ (see figure at right); in the bridge rectifier $v_{out} = |v_{in}| - 2V_f$. In both of these expressions, if the right-hand side ≤ 0, then $v_{out} = 0$: a diode remains off until the voltage drop across it has the correct polarity and reaches V_f. Similarly, nonzero I_R will flow in the opposite direction through a diode when it is reversed biased; this implies that v_{out} can become very slightly negative: $v_{out} = -R_{load} I_r$.

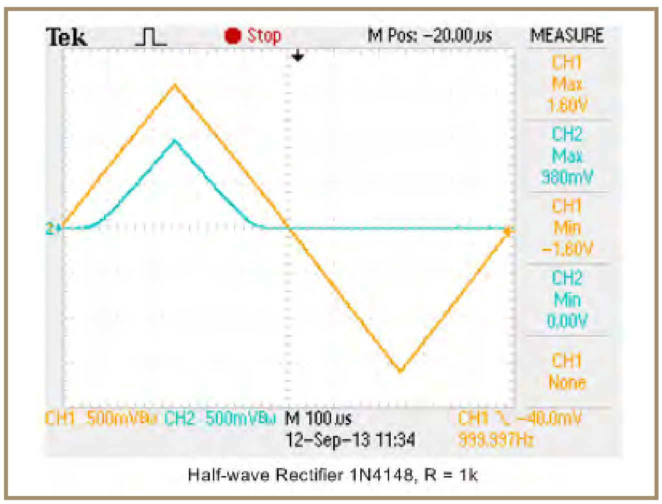


Figure 3-3: Response of the half-wave rectifier showing the effect of V_f, the diode's forward voltage drop. The diode is silicon, so V_f ≈ 0.6V (input is CH1, output CH2).

Precision rectifier circuits

A diode's forward voltage drop of a few tenths of a volt (Figure 3-3) means that the basic diode circuits of the previous section can hardly be considered to provide precision rectification of an input signal, especially if the signal is small. As you might expect, adding the capabilities of an operational amplifier may be able to remedy this situation: consider the circuit in Figure 3-4.

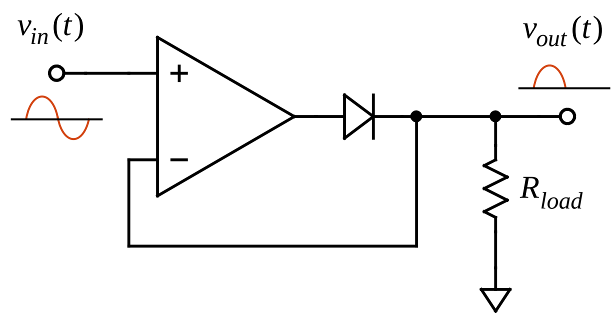


Figure 3-4: Simple, precision half-wave rectifier. Placing the diode inside the feedback loop ensures that the op-amp output will compensate for the diode's forward voltage drop. When v_{in} < 0, the diode turns off as the op-amp output goes negative; R_{load} then ensures that v_{out} = 0 in this case.

This circuit is essentially a voltage follower, but a diode has been added in series with the op-amp output *inside the feedback loop*. When $v_{in} > 0$, the op-amp will set its output one diode drop higher than v_{in} so that its *–Input* will equal v_{in}. Thus, for $v_{in} > 0$, $v_{out} = v_{in}$ even though the diode may have a significant forward voltage drop. When $v_{in} < 0$, the op-amp's output will go negative, and, because the diode's cathode is connected to ground through R_{load}, the diode will turn off, disconnecting the op-amp's output from v_{out}. Thus, for $v_{in} \leq 0$, $v_{out} = 0$. The load resistor value should be around 1kΩ; if low output impedance is required, buffer the output with a voltage follower.

Adding another op-amp, we can construct a precision full-wave rectifier, Figure 3-5. The second amplifier (using op-amp U2) combines the original input signal with the half-wave rectifier output to form a fully rectified waveform (the absolute value of the input waveform).

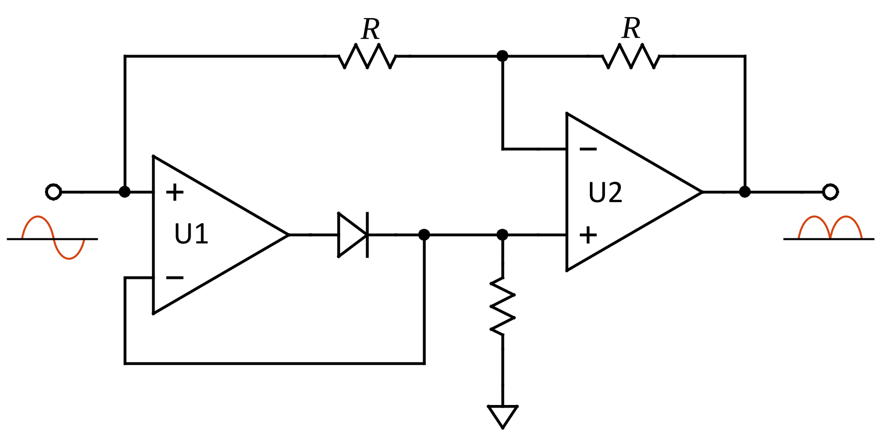


Figure 3-5: Precision full-wave rectifier. The output of a half-wave rectifier is combined with the original input signal by the amplifier U2. Its output is the absolute value of the input. The two resistors R should be well-matched in value (1% or better).

The two resistors used to form the feedback network for U2 should be well-matched in value. Note that U2 is configured as the combination inverting+noninverting amplifier of Experiment 1 Figure 1-19 on page 1-22, with its v_{in+} coming from the output of the half-wave rectifier, U1. Analysis of this circuit is left to the exercises. Typical outputs from the two rectifier circuits are shown in Figure 3-6; note that the effect of the forward diode drop has been eliminated (compare with Figure 3-3 on page 3-4).

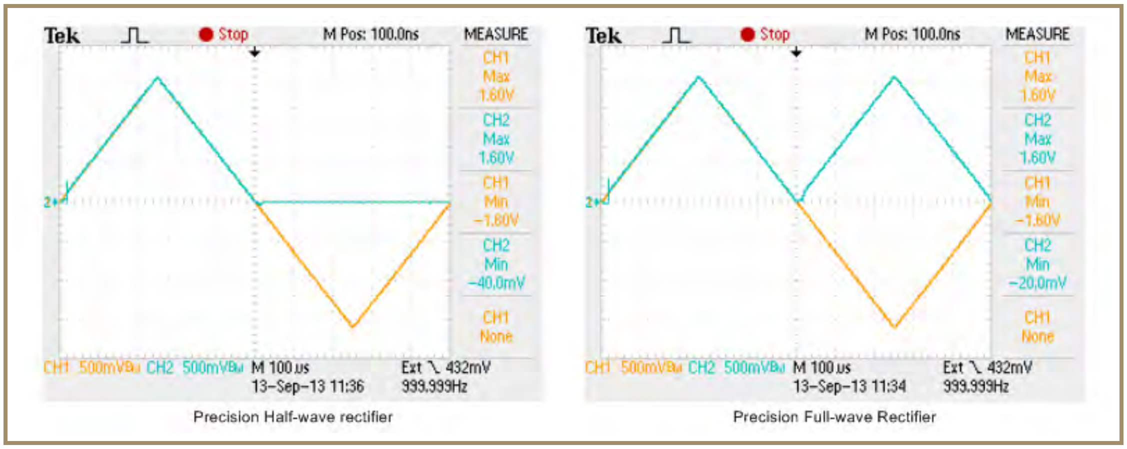


Figure 3-6: Precision rectifier outputs. Left: half-wave rectifier circuit of Figure 3-4. Right: full-wave rectifier of Figure 3-5. Input signal waveform and display scales are the same as in Figure 3-3 on page 3-4; note that the precision rectifier circuits eliminate the 0.6V forward voltage drop of the silicon diode used in the circuits. The small "glitches" in the output waveforms as the input goes through 0 are effects of the TL082 op-amp's finite slew rate.

Before proceeding further, some important diode limitations should be noted. For the 1N4148 silicon small-signal diodes (the type you will mostly use) the limits are:

$$V_R = 75\,\mathrm{V} \qquad I_F = 300\,\mathrm{mA} \qquad P_D = 0.3\,\mathrm{W}$$

DIODE LIMITATIONS

Exceeding these limits could cause catastrophic diode and circuit failure:

V_R ***reverse breakdown voltage***: the maximum reverse-bias voltage which may be safely applied without the diode exhibiting avalanche or Zener breakdown.

I_F ***maximum forward current***: the maximum current the diode can tolerate when forward-biased.

P_D ***maximum power dissipation***: the maximum power dissipation the diode can tolerate without overheating and failing.

Warning

Because the forward voltage drop, V_f, of a diode is nearly independent of the current flowing through it, *the magnitude of the forward current must be limited by the external circuit*, or a forward-biased diode will quickly fail.

Exponential and logarithmic amplifiers

The exponential voltage-current relationship of the PN junction diode (equation 3.9 on page 3-32) may be exploited to build amplifiers with approximately exponential or logarithmic gains. If the diode current is much greater than I_R in equation 3.9, then the I-V relationship is approximately:

$$I = I_R e^{q_e V/\eta k_B T}$$

where, for a silicon diode, $I_R \sim 10^{-6}\text{mA}$ and $q_e/\eta k_B T \approx 20\text{V}^{-1}$ at room temperature. Thus, V will change by approximately 0.12V for a factor of 10 change in the current I, and $V \approx 0.6\text{V}$ for $I = 1\text{mA}$. Thus we can take the logarithm or the antilog (exponential) of an input voltage using the circuits in Figure 3-7. With resistor value R, the transfer functions are approximately:

$$\left|V_{out}\right| \approx 0.6\text{V} + 0.12\text{V} \times \log_{10}\left|\frac{V_{in}/\text{Volt}}{R/\text{k}\Omega}\right| \quad \text{(log amplifier)}$$

3.1

$$\left|V_{out}\right| \approx (1\text{mA} \times R)10^{\left|V_{in}/0.12\text{V}\right| - 5.0} \quad \text{(exponential amplifier)}$$

The sign of the output voltage in each case is the opposite of the sign of the input, since the amplifiers are inverting. The gains and offsets of these circuits may be inconvenient, so more circuitry is usually added to scale and offset the output signals; the transfer function, however, is quite temperature-dependent, which lets us smoothly segue to the next section.

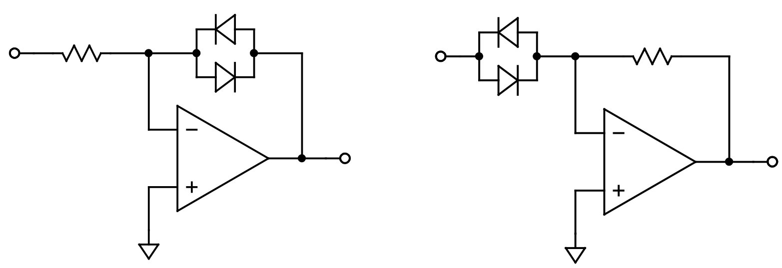


Figure 3-7: Inverting log (left) and exponential (right) circuits using diodes. These simple circuits are not very accurate and are very temperature-sensitive, but will work for noncritical applications. The paralleled back-to-back diodes will treat positive and negative input voltages the same, using whichever diode is forward-biased to dominate the circuit transfer function.

Temperature sensing

The PN junction's temperature sensitivity (equation 3.9 on page 3-32) makes the semiconductor diode a useful temperature sensor; in this section we present one example of a temperature monitor circuit. If a silicon diode is forward-biased so that its current is much greater than I_R in equation 3.9, then the I-V relationship is:

3.2
$$I \approx I_0 \exp\left[\frac{-q_e}{\eta k_B}\left(\frac{V_g - V}{T}\right)\right]$$

This implies that if a diode forward-bias current is held constant at, say, 0.1mA ($\sim 10^5 I_R$), then $V_g - V \propto T$. For a silicon diode at room temperature, $(V_g - V)/T \approx 2\,\mathrm{mV/K}$, and this is the sensitivity of the forward-bias voltage to a change in junction temperature. Here is a circuit:

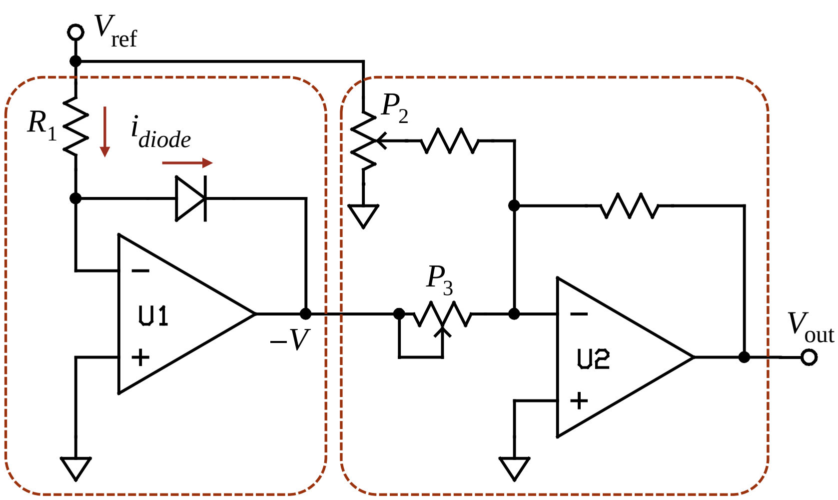


Figure 3-8: Temperature circuit using a diode sensor. Amplifier U1 maintains a constant diode forward-bias current of V_{ref}/R_1; the amplifier output is then $-V$, where V is the diode's forward-bias voltage. The inverting, summing amplifier U2 scales and offsets this voltage to provide the circuit's output (potentiometer P_2 trims the output voltage's DC offset, P_3 trims $\Delta V_{out} / \Delta T$).

To provide an output temperature response of $0.1\,\mathrm{V/°C}$, the gain of the U2 inverting amplifier stage should be ≈ 50 when R_3 is set to its center position; the summing amplifier's gain for the offset adjust input (R_2) should be 1 or less, so when R_2 is set near its center position the output corresponds to the diode temperature. There are several ways this circuit could be refined, so that constructing and calibrating such a circuit could be a good final project. The Texas Instruments LM35 series of analog temperature sensor ICs implements a version of the circuit in Figure 3-8 in a single, small, calibrated device.

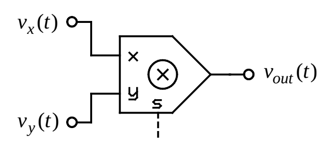
THE ANALOG MULTIPLIER

The ideal analog multiplier

Unlike the simple semiconductor diode, the modern analog multiplier is a sophisticated, complicated integrated circuit incorporating temperature-compensated voltage references, matched transistors, and laser-trimmed resistors. Much like the modern operational amplifier, this internal circuitry makes these devices particularly simple to utilize (although actual multipliers tend to behave in a less ideal manner than their op-amp cousins). In this section we consider the properties of an ideal analog multiplier and discuss applications of such a device; the next section will look at the use and limitations of an actual multiplier IC.

An analog multiplier, of course, requires at least two user-controlled input signals, which are typically called the x and y inputs. The input and output analog values may be voltages or currents, depending on the particular device, but modern medium-speed, precision multipliers usually input and output *voltages*, so that is what we will assume here. The product of two voltages has, of course, units of Volts^2, but the output will be in Volts. Thus the product xy is divided by an additional voltage parameter, the *scale factor*, s, which converts the product from Volts^2 to Volts (equation 3.3).

3.3 **Multiplier:** $v_{out} = v_x v_y / s$



The analog multiplier usually has a precision, built-in scale factor which is often $s = 10.0\text{V}$, but it may include an external terminal so that you can adjust the value of s (as shown in the schematic above). However the scale factor may be set, it is usually limited to positive values $(s > 0)$. Modern analog multipliers don't limit the signs of v_x and v_y (they support "four-quadrant multiplication"), but they usually limit their magnitudes to $\leq s$.

Placing a multiplier in an op-amp's feedback loop provides for the calculation of the inverse operation, division, as shown in Figure 3-9. Assuming the negative feedback around the op-

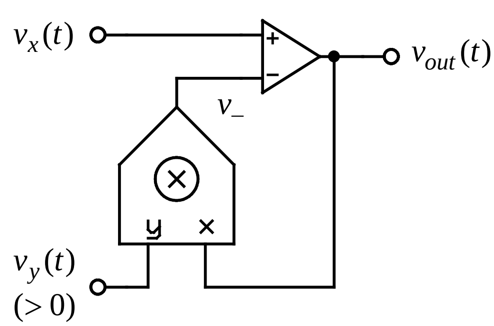


Figure 3-9: Basic divider circuit using an op-amp. Note that for the negative feedback to be effective, v_- must have the same sign as v_{out}, and this condition requires that $v_y > 0$. Improving this circuit to allow "four-quadrant division" so that $v_y < 0$ is supported is a nontrivial exercise and could make a good final project.

amp is effective, then $v_- = v_+ = v_x$. But the multiplier output is $v_- = v_{out} v_y / s$, so the divider circuit output is:

$$\text{3.4} \qquad \textbf{Divider:} \quad v_{out} = s v_x / v_y \qquad (v_y > 0)$$

Since a typical multiplier will require that its inputs not exceed *s*, the gain of the circuit in Figure 3-9 with respect to the op-amp's *+Input* (v_x) will be greater than 1, and thus the circuit's bandwidth will be less than the op-amp's gain-bandwidth product f_{BW}; in fact, the bandwidth will be $f_{BW}(v_y/s)$.

> The condition $v_y > 0$ in (3.4) comes from the requirement that v_- must have the same sign as v_{out} (see Figure 3-9), or the feedback will effectively change sign and become *positive*, which will cause the op-amp's output to proceed to one of its power supply limits (i.e. the output will *saturate*) until v_y becomes positive.

You may calculate the square of a signal by connecting a multiplier's inputs together; the inverse operation, a square root circuit, again may be built by putting the multiplier in an op-amp feedback loop, but now circuit *latch-up* is a real possibility: the op-amp output will saturate (in this case at its negative limit) and *stay there* no matter what the input does. Using a precision rectifier configuration (as in Figure 3-4) is a popular way to avoid this latch-up problem (Figure 3-10).

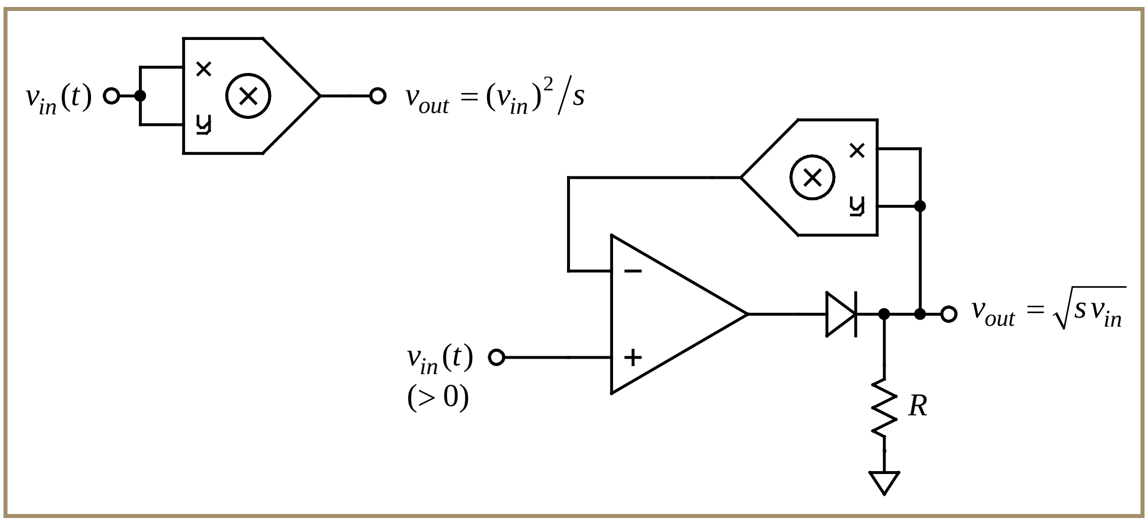


Figure 3-10: Square and square root circuits. The square root circuit is prone to *latch-up*: if the inputs to the multiplier were to go < 0, even momentarily, its positive voltage output would drive the op-amp into saturation at its negative voltage limit, permanently maintaining this undesirable state. The diode+resistor combination on the op-amp output ensures that the multiplier inputs never go negative, avoiding latch-up. A value of about 1kΩ – 10kΩ is appropriate for R. Of course, s is the multiplier's *scale factor*.

A real analog multiplier IC

Because circuit designers often need to include op-amps and need to invert signals in their circuits, actual IC multiplier devices usually include extra circuitry to make the designer's job easier. In this section, we consider the Texas Instruments MPY634 device, which is included on the *ASLK PRO* breadboard; a similar, general purpose, relatively inexpensive multiplier is the Analog Devices AD633, which may be more readily available.[1] In this section we discuss the use of the quite complicated MPY634, a functional block diagram of which is shown below.

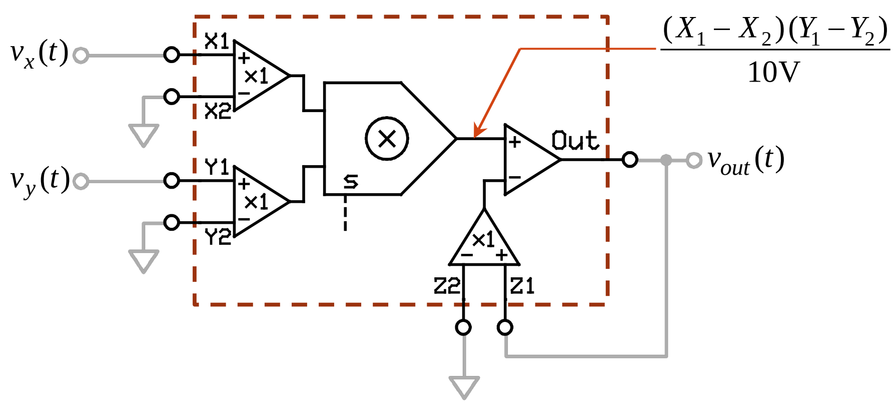


Figure 3-11: Functional block diagram of the Texas Instruments MPY634 analog multiplier. Besides the basic multiplier, the IC includes a general-purpose op-amp on the multiplier's output and has differential inputs for all parameters, including the *Z* inputs to the op-amp's feedback (*–Input*) terminal. Shown in light gray are the connections needed to emulate the basic, generic multiplier discussed in the previous section. Note that the ×1 amplifiers on the multiplier inputs *are not op-amps* – they simply apply the difference of their two inputs to their outputs (X1–X2, etc.).

The MPY634 has differential inputs for all input parameters and includes a general-purpose op-amp on the multiplier output so that it can be easily configured for different scale factors or for inverse operations. In Figure 3-11 the output op-amp has been configured as a voltage follower (output fed back through its Z1 input), and the negative differential input for each parameter has been grounded; the resulting circuit acts as the basic multiplier described by equation 3.3 in the previous section. Note that the default scale factor is 10V; this is the value if the IC's scale factor terminal (**S** in the figure) is left disconnected. The MPY634 basic multiplier accuracy in this configuration is approximately 2%. The op-amp $f_{BW} \geq 6\text{MHz}$, and it has a $20\text{V}/\mu\text{sec}$ slew rate, so it has similar performance to the TL082 op-amps on the breadboard.

Divider and square root configurations are shown in Figure 3-12 (on page 3-12). Because the multiplier output is connected to the internal op-amp's *+Input*, you have to *invert the multiplier's output* to provide negative feedback around the op-amp (as in Figure 3-9 and

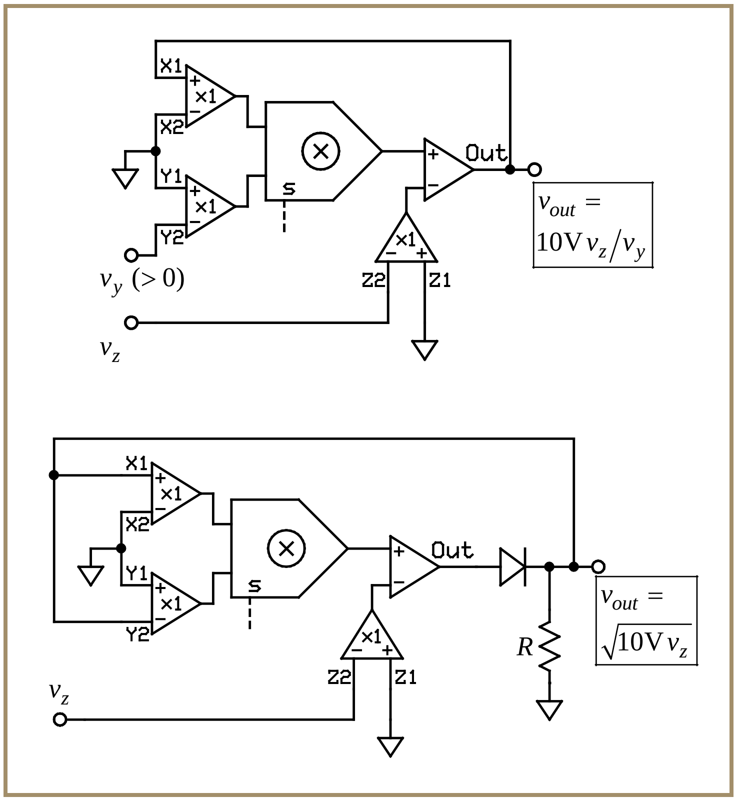
[1] "Inexpensive" is a relative term. The TL082 IC you've used costs less than $1 and includes two complete op-amp devices, whereas the AD633 IC cost nearly $11 for a single multiplier, and the MPY634 on the *ASLK PRO* board costs nearly $25 each (prices as of fall 2017).

Figure 3-10). This is accomplished by using the (–) differential terminal for one of the arguments to the multiplier, so its output is the negative of the product of its two inputs. Similarly, the other op-amp input (to its –*Input* terminal) is also inverted by using the input's (–) differential terminal. The MPY634 data sheet has more examples demonstrating this device's flexibility.

Figure 3-12: MPY634 divider (top) and square root (bottom) circuits.

In each case the multiplier sub-circuit must be in the *negative feedback loop* of the op-amp (see Figure 3-9 and Figure 3-10); to accomplish this using the MPY634, the multiplier output must be the *negative* of the product of its inputs, which is accomplished by *inverting only one* of its inputs. The v_z input to the op-amp is also inverted, so that it effectively becomes a *noninverting* op-amp input (−1 × −1 = +1).

A diode and resistor are still needed to prevent latch-up of the square root circuit output, as discussed in the previous section. The resistor value R should be about 1kΩ – 10kΩ, as before (Figure 3-10).



You must always provide negative feedback for the output op-amp of the MPY634.
Any unused differential input terminals must be connected to ground.

PRELAB EXERCISES

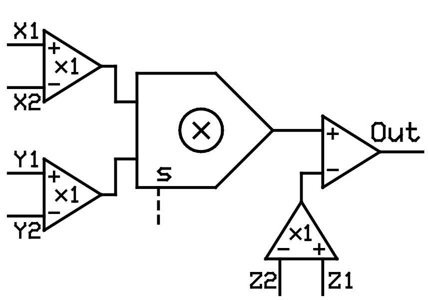
1. Consider the simple half-wave rectifier circuit (left-hand schematic in Figure 3-2 on page 3-3). If the diode is perfect (both its forward voltage and reverse leakage current are 0), then what is the circuit's *output resistance* during a positive half-cycle (when $v_{in}(t) > 0$)? What about during a negative half-cycle ($v_{in}(t) < 0$)? Include R_{load} as part of the circuit when calculating the output resistance.

   Hint: review the section on *output resistance* starting on page 1-43 of Experiment 1. Consider two distinct cases: one when the diode is forward-biased (and, since $V_f \equiv 0$, its resistance = 0), and the other when the diode is reverse-biased (so, since $I_R \equiv 0$, its resistance is infinite).

2. How does the full-wave rectifier circuit in Figure 3-5 (on page 3-5) work? Analyze the two circuit operating states separately: (1) the diode is forward-biased; (2) the diode is reverse biased. Assume the op-amps are ideal; for each of these states determine:
   a. the required input voltage (v_{in}) range for the circuit to be in that state
   b. the voltage at U2's +Input
   c. the circuit's transfer function v_{out}/v_{in}

3. Consider the square root circuit in Figure 3-10 on page 3-10. If the input $v_{in}(t) < 0$, then what is v_{out}? What is the output voltage of the op-amp? What happens to the op-amp output and v_{out} as the input rises through 0? Does the forward-bias diode voltage drop affect the accuracy of v_{out} (assuming the op-amp is ideal)? Why or why not?

4. By configuring the MPY634's output op-amp to provide gain, you can effectively adjust the device's multiplication scale factor to something less than its 10V default. Complete the schematic diagram below by adding a feedback network to the op-amp and properly connecting the input terminals so that you have a multiplier with a scale factor $s = 5\text{V}$ (assign appropriate values to any resistors you include).



Another problem on the next page…

5. Design a circuit using the MPY634 to create an amplifier with a *voltage-controlled gain*. The gain control input voltage range should be −7V to +7V (minimum), and the circuit gain should equal the control voltage in volts, e.g. −2V → gain = −2, etc. The allowable signal input and output voltage ranges (without *clipping* or distortion of the output) should be −7V to +7V as well (as long as the circuit gain is not set too high). Use additional op-amp amplifier stages as part of the circuit design if you need them (but you may not need them). Assume that the input signal is ground-referenced, so you don't need a fully differential input for the signal. The gain accuracy should be no worse than ±10%.

   Provide a full schematic of the circuit with all component values included. Use only these standard 5% tolerance resistor values: a power of 10 times…

   1.0, 1.1, 1.2, 1.3, 1.5, 1.6, 1.8, 2.0, 2.2, 2.4, 2.7, 3.0, 3.3, 3.6, 3.9, 4.3, 4.7, 5.1, 5.6, 6.2, 6.8, 7.5, 8.2, 9.1

   You will build and test this circuit during your lab session, so think carefully about it!

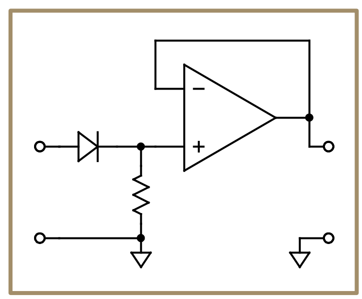
LAB PROCEDURE

Overview

The semiconductor diode and the analog multiplier are two examples of nonlinear components, but they are nevertheless very different in their behaviors and their applications. Spend sufficient time during lab building circuits with each of them so that you become comfortable with using such elements in your designs. Spend some time investigating light emitting diodes (LEDs) and Zener diodes as well as the basic silicon diode. Combine circuits using them with the amplifier and filter designs you have already practiced with so that you start to think about and develop more complicated applications.

Detailed procedures

Diode rectifiers and an AM demodulator



Build a simple half-wave rectifier circuit using a 10k load resistor and a voltage follower to buffer the rectifier's output (figure at right). Using a 1kHz signal input of a couple of volts peak-peak amplitude, confirm that the diode's turn-on voltage is approximately 0.6V, as shown in Figure 3-3. Replace the diode with an LED. How to determine the LED polarity (anode v. cathode) is shown in Figure 3-13. Note that you must increase the signal generator output to at least 3V peak-peak to get the LED to conduct. What is its approximate turn-on voltage?

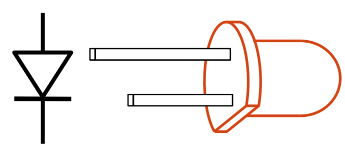


Figure 3-13: Determining LED polarity. The long lead is the anode (+); the short lead (nearest the flat section of the LED circumference) is the cathode (–).

Again using a silicon diode, replace the simple half-wave rectifier circuit with a precision version (Figure 3-4). Examine the output and note that the 0.6V diode drop is no longer evident. Compare your results to Figure 3-6.

Now filter the rectified output by connecting first the 0.1μF and then the 1μF capacitor in parallel with the 10k load resistor. Compare your results to the filtered half-wave rectifier output plotted in Figure 3-14 on page 3-17.

Using the 0.1μF capacitor to filter the rectifier output, next configure the signal generator to input an *amplitude modulated* (AM) signal to your rectifier. Use a *carrier frequency* of 100kHz, a *modulation frequency* of 200Hz, and a AM *modulation depth* of 30%. Get your TA to help you set up the signal generator; trigger the oscilloscope using the rectifier circuit output. Does the rectifier output *detect* the modulation (output mostly the 200Hz *envelope* of the 100kHz carrier)? Try different load resistor and filter capacitor combinations. Try varying the modulation frequency and modulation depth and note the effects on the output.

Using a multiplier as a frequency doubler

Construct a basic squarer circuit using one of the analog trainer's MPY634 multipliers; refer to Figure 3-11 on page 3-11 for the basic multiplier connections. Using a 5V peak-peak input signal, confirm that the squarer's output is given by $v_{out}(t) = v_{in}(t)^2 \big/ 10\text{V}$.

Once the circuit is working and behaves as expected, note that the circuit's response to a sinusoid input is, of course, another sinusoid at twice the input frequency + a constant (DC) offset (since, of course, the output of the squarer is nonnegative). You may remove the DC component by *AC coupling* the output, as you learned in Experiment 2. Add an RC high-pass filter to couple the squarer output to a noninverting, gain 11 amplifier stage.

You have now constructed a *frequency doubler* — a sinusoid input produces a sinusoid output at twice the frequency. Using an input frequency of approximately 1kHz, determine what input signal amplitude is required to produce an output of the same amplitude. Increase the input frequency until you find the −3dB upper bandwidth limit of your doubler.

Amplifier with voltage-controlled gain

Build and test the circuit you designed for prelab exercise 5. The gain control voltage may be generated by a computer *DAQ* analog output port which you control using the National Instruments *Measurement & Automation Explorer* application; the lab instructor or your TA will show you how to accomplish this.

Additional, self-directed investigations

Maybe try building a cubing circuit: $v_{out} \propto v_{in}^3$; how many multipliers would this take? Consider a log or exponential amplifier, or one or more circuits from the **MORE CIRCUIT IDEAS** section (such as the true RMS circuit), or try a nontrivial one of your own design. Look back at earlier experiments to see if there are any other circuits you would like to investigate.

Lab results write-up

As always, include a sketch of the schematic with component values for each circuit you investigate, along with appropriate oscilloscope screen shots and, if appropriate, Bode plots. Make sure you've answered each of the questions posed in the ***Detailed procedures*** section.

More circuit ideas

Peak detectors

If you put a capacitor in parallel with the output of a simple, half-wave rectifier as shown in Figure 3-14 below, then whenever the diode conducts the AC voltage source will quickly charge the capacitor up to match its voltage (minus the diode's forward voltage drop). When the AC voltage source passes its peak positive output and starts to decrease, its voltage drops below the capacitor's voltage, and the diode turns off. The capacitor then begins to relatively slowly discharge through the load resistor until the AC source voltage output again becomes high enough to turn the diode back on and recharge the capacitor. As a result, the output voltage stays near the peak positive input voltage value; the smaller the capacitor or the smaller the load resistor, then the more the capacitor will discharge during a cycle. This simple circuit is useful as part of a *power supply* to turn the 60Hz power line AC voltage (usually coupled through a transformer) into a nearly constant DC voltage to power an electronic device. The amplitude of the capacitor charge-discharge output oscillation is called the power supply *ripple voltage*; small ripple voltage is desirable, so power supply filter capacitors tend to be large; they are usually electrolytic (refer again to the photo in Experiment 2, Figure 2-1 on page 2-6).

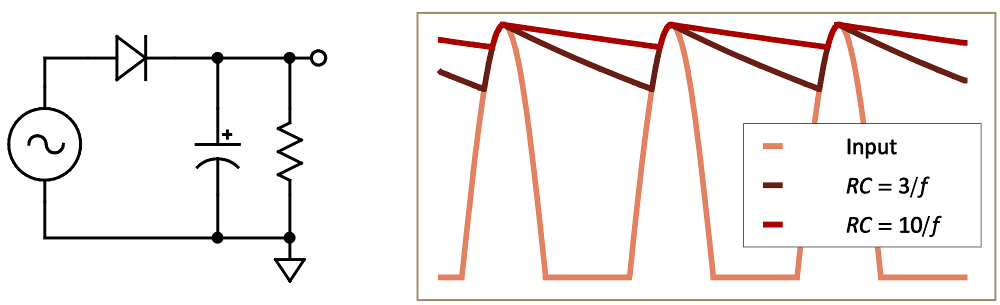


Figure 3-14: Using a filter capacitor to smooth the output of a simple, half-wave rectifier. The capacitor C is charged whenever the input voltage exceeds the output voltage so that the diode is forward-biased; when the input voltage drops and the diode becomes reverse-biased, the capacitor discharges through the load (the resistor R). The graph at right shows the resulting output voltage variation for various RC time constant values; f is the frequency of the input source. The effect of the diode's forward voltage drop is not included in the graph.

If we use this idea with a precision half-wave rectifier circuit, we get the *peak detector* shown in Figure 3-15, which works just the same as the simple circuit discussed above, except now an op-amp is used to correct for the diode forward voltage drop, and a voltage follower stage is added to the output so that the subsequent load will not discharge the capacitor. This circuit will hold the maximum positive input voltage encountered for quite a long time; if a resistor is not included and a high-quality capacitor is used, then the capacitor's discharge will only be because of the follower op-amp's input bias current and the diode's reverse leakage current, the sum of which can be kept to only a few nanoamps by careful component selection. In this case, a 10μF capacitor would provide an output voltage which decayed at less than $1\text{mV}/\text{sec}$. If you need to hold a peak voltage value for longer than this, then your best bet would be to record a *digitized* version of the output voltage value.

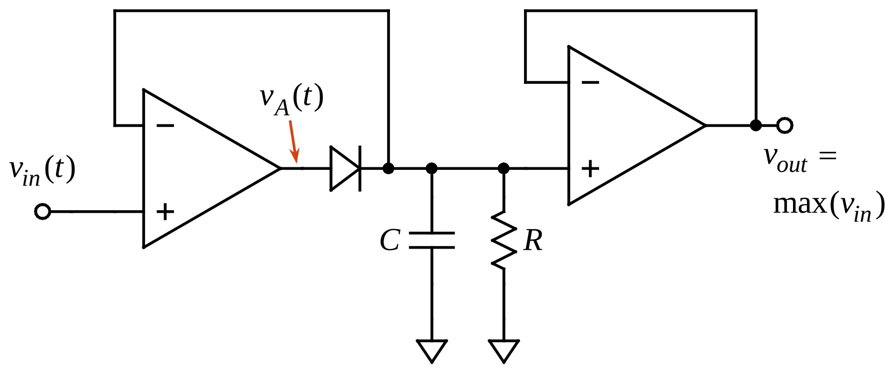


Figure 3-15: A peak detector circuit, which outputs the most positive value of the input signal seen so far; reversing the diode would output the most negative value of the input. The output voltage will exponentially decay toward 0 with time constant RC; because of the voltage follower on the output, RC may be made quite long by choosing a large value for R, or the resistor could even be replaced by a switch to reset the output to 0. The behavior of the first op-amp's output voltage, v_A, is discussed in the text.

One potential problem with the peak detector shown above is that it can output a voltage significantly lower than the peak amplitude of a high frequency input signal. This is a big problem for some applications, so now consider ways to improve the peak detector's speed. There are two major design issues which increase the time it takes to change the capacitor's voltage: (1) the amount of current the op-amp output can supply, and (2) the op-amp slew rate. The first issue is easy to deal with, so we consider it first. When the input exceeds the voltage currently stored by the capacitor, the diode becomes forward-biased, and the op-amp output can charge the capacitor toward the new maximum voltage. The rate that the capacitor's voltage will change is given by $i_{max} = C\, dv/dt$, where i_{max} is the maximum output current the op-amp is capable of supplying (usually specified in the op-amp's data sheet). The TL082, for example, can output up to about 40mA into a discharged capacitor, but its available output current decreases as its output voltage rises. If your application must track a signal whose peak amplitude changes rapidly, use as small a capacitor value C as you can (consistent with your required peak hold time) or get an op-amp with a larger current output capacity.

The second problem is more difficult to handle, because the design of the circuit in Figure 3-15 exacerbates the speed limitation imposed by the op-amp slew rate. Whenever the input is less than the voltage stored by the capacitor, the op-amp output voltage (v_A in Figure 3-15) goes all the way down to its negative limit, because the diode is reverse-biased (opening the feedback loop), and therefore $(v_{in} =) v_+ < v_- (= v_{out})$. If and when the input returns to above

v_{out} (so that $v_+ > v_-$), the op-amp output must change from its negative limit up to a diode-drop above v_{out} in order to forward-bias the diode and start to increase the capacitor's charge. The time it takes to change the output voltage is, of course, limited by the op-amp slew rate. Since the output starts at the op-amp's negative rail, the output must change by at least several volts, which will take about a microsecond or more for the TL082 ($13\,\mathrm{V}/\mu\mathrm{sec}$ slew rate). One solution, obviously, would be to get a faster op-amp (the TI THS3201-EP, for example, has $9800\,\mathrm{V}/\mu\mathrm{sec}$ slew rate, 1.8GHz f_{BW}, 100mA output current, and costs \$5.54 each) — this is the only solution if your circuit must respond to very narrow, infrequent pulses. If, on the other hand, you must track the amplitude of a high-frequency sinusoid, a simple modification to the design of the peak detector circuit can greatly increase its frequency response (Figure 3-16).

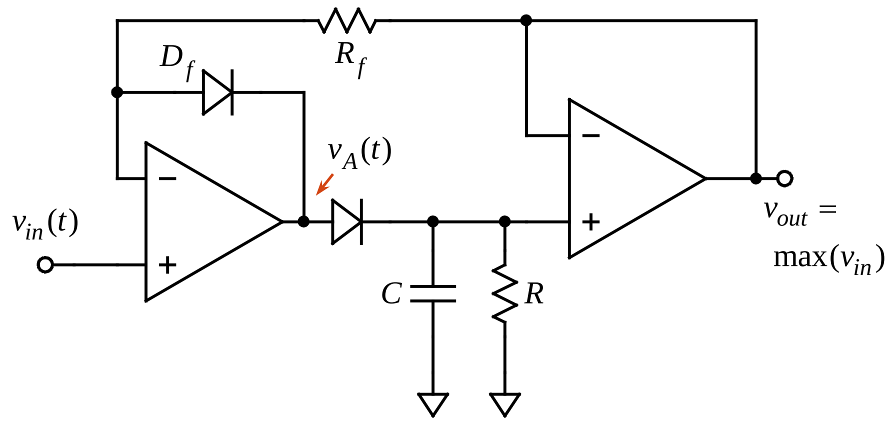


Figure 3-16: A modified peak detector with improved frequency response. The feedback loop now comes from the voltage follower output; the addition of R_f and D_f ensures that the loop stays closed when v_{in} drops below v_{out}. Now the first op-amp output (v_A) never gets further than a diode drop away from v_{in} while it is below v_{out}, rather than saturating at its negative limit as is the case in the Figure 3-15 design.

This modified design adds resistor R_f and diode D_f. First consider the state when v_{in} has been rising through the previously stored peak voltage, so that the output diode is conducting and C is being charged. Now $v_{out} = v_C = v_- = v_{in}$, and v_A is a diode drop higher, so the feedback diode D_f is reverse-biased and doesn't affect the circuit's operation (no current flows through R_f in this case, ensuring that $v_- = v_{out}$). When v_{in} passes its new peak and starts to drop, the op-amp output v_A rapidly decreases, turning off the output diode. Momentarily, $v_- = v_{out}$ stays constant, but once v_A has fallen a diode drop below v_- the feedback diode starts to conduct, and v_- will now follow v_A, remaining a diode drop above it. Since $v_{out} = v_C$, v_{out} remains constant, and current flows from the output op-amp through R_f and D_f to the first op-amp output (at voltage v_A). The first op-amp output voltage will slew down until $v_- = v_{in}$, and then will maintain that condition by keeping v_A a diode drop below v_{in}.

As v_{in} again rises and approaches the stored voltage v_{out}, v_A is only one diode drop away (unless the rate of change of v_{in} exceeds the op-amp slew rate). Thus as v_{in} passes v_{out}, the op-amp output only has to slew by two diode drops to turn on the output diode and start charging C, rather than slewing all the way up from its negative output rail. A two-diode-drop slew takes only about 0.1μsec for the TL082, an order of magnitude quicker than a slew of 13V up from its saturated negative output (at about −11V for our breadboards).

Zener diode regulator

Some diodes are designed to be used in their *reverse-bias breakdown region*: *Zener* and *avalanche* diodes. These types of diodes are very useful as voltage references, simple voltage regulators, and overvoltage protection devices. The Zener and avalanche breakdown effects are described very briefly in the section **THE PN JUNCTION DIODE** on page 3-26; in this section we address one common application of these devices.

Figure 3-17 shows a typical Zener reverse-bias *I-V characteristic curve* — a plot of the relationship between applied reverse-bias voltage and resulting current flow through the diode. The very steep portion of the curve corresponds to the diode's reverse-bias breakdown region; because the curve in this region is so steep, you can see that changes in the diode reverse current correspond to very small changes in reverse-bias voltage. Thus, *the voltage across the Zener diode in this breakdown region is very insensitive to changes in the current through it*. This characteristic makes the Zener diode useful as a simple *voltage regulator*.

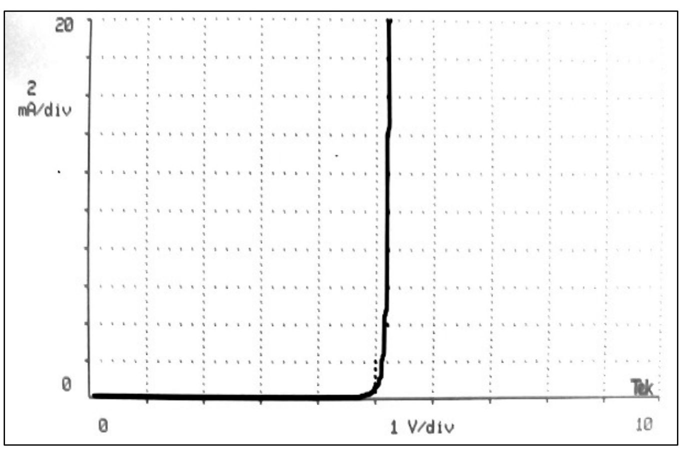


Figure 3-17: Measured Zener diode I-V curve. The reverse diode current is plotted as a function of the applied reverse-bias voltage. As the applied voltage exceeds 5V, the diode current dramatically increases as the diode suffers reverse breakdown. As can be seen from the plot, diode reverse currents above about 15 mA correspond to a reverse-bias voltage of 5.3V. A lab instrument called a *curve tracer* was used to perform this measurement (Tektronix 571).

Assume we have a power supply whose output voltage is greater than the maximum voltage allowed by some device you need to power, and this device needs a *well-regulated power supply voltage* (the voltage is largely unaffected by changes in load current or the input voltage supplied to it). By using a Zener diode whose reverse breakdown voltage is the same as the voltage you need to power your device, you can build a simple voltage divider circuit which will satisfy the requirements for your device's power supply.

Consider the circuit in Figure 3-18, which is just a voltage divider with a reverse-biased Zener diode as the bottom element (note how the cathode end of the schematic symbol for a Zener diode differs from that for a regular diode). If the input voltage exceeds the Zener diode's breakdown voltage, then the diode may break down, with voltage V_R across it. Thus $v_{out} = V_R$. Now the voltage across the resistor R is determined: $v_{in} - V_R$, and the current through the resistor is also known: $i_{in} = (v_{in} - V_R)/R$. This current is divided between the current through the Zener diode, i_Z, and that through the load, i_{load}. Thus $i_Z = i_{in} - i_{load}$.

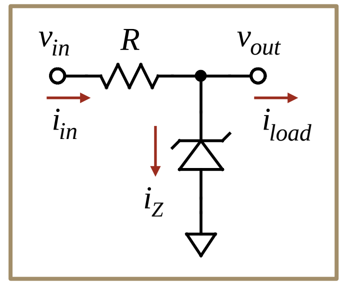


Figure 3-18: Zener diode voltage regulator circuit. The input voltage v_{in} must be greater than the Zener diode's reverse breakdown voltage, V_R. This large voltage causes the diode to suffer reverse breakdown, with the difference v_{in} – V_R imposed across resistor R, establishing the input current i_{in}. This current is divided between that required by the load, i_{load}, and the reverse current through the diode, i_Z.

Note how this circuit holds the voltage applied to the load constant (equal to V_R) even if i_{load} or v_{in} varies. Changing i_{load} doesn't change v_{out} or i_{in} because the voltages across the diode and across R don't change; so for changes in i_{load}, $\Delta i_Z = -\Delta i_{load}$ (we assume here that the Zener's I-V characteristic curve is very steep in its breakdown region: $dV_R/di_Z \approx 0$). Similarly, changes in v_{in} don't affect v_{out} because, although i_{in} changes, $\Delta i_Z = \Delta i_{in} = \Delta v_{in}/R$, but the steep Zener I-V characteristic ensures that v_{out} is nearly unaffected by this change in i_Z.

DESIGNING A ZENER VOLTAGE REGULATOR CIRCUIT

This example will illustrate how you would design a voltage regulator using the Zener diode circuit in Figure 3-18. Assume the load is a digital circuit which requires a stable voltage of no more than 5.5V in order to operate (the so-called CMOS *digital logic family* to be discussed in a later experiment would fall into this category), and you wish to power it from a 9V battery. The circuit will require a minimum of 10mA, but could draw as much as 20mA when an LED (light emitting diode), a part of the circuit, is illuminated. You want the circuit to work properly even if the battery has discharged to the point where it can only supply 7V.

Here are the basic Zener regulator design steps:

1. Choose a Zener diode breakdown voltage. The Zener diode whose characteristic curve is shown in Figure 3-17 should work, since its nominal breakdown voltage of 5.3V is near, but less than, the specified 5.5V limit.

2. Choose a *minimum* Zener current which will give good voltage regulation (steep I-V curve); according to Figure 3-17, a current of 10mA is well into the diode's breakdown region, so we will design the circuit for that minimum diode current.

3. The minimum Zener current + the maximum load current = the design target for the current i_{in} through the resistor R (Figure 3-18). Therefore, for this example, design $i_{in} = 10\text{mA} + 20\text{mA} = 30\text{mA}$. The resistor value is then chosen to give the correct voltage drop at this design current when the input voltage is at its *minimum* (7V for this example). Thus $R = (7\text{V} - 5.3\text{V})/30\text{mA} = 56.7\Omega$. Choose the closest standard resistor value, which, in this case, is $\boxed{R = 56\Omega}$.

4. Now consider what would happen when the circuit experiences the opposite extreme: maximum source voltage v_{in} (9V) along with minimum load current i_{load} (10mA). Since the Zener diode will keep the output voltage at 5.3V, the voltage drop across R is now $9\text{V} - 5.3\text{V} = 3.7\text{V}$, and with the chosen value for R, $i_{in} = 3.7\text{V}/56\Omega = 66\text{mA}$. Under these conditions the diode current will have increased to 56mA ($i_Z = i_{in} - i_{load}$).

5. Use the results from (4) to determine the worst-case power dissipations in the resistor and the diode:

   Resistor: $P = VI = 3.7\text{V} \times 66\text{mA} = 0.24\text{W}$
   Diode: $P = VI = 5.3\text{V} \times 56\text{mA} = 0.3\text{W}$

   These results specify the required minimum power dissipation capabilities of the components, which should be at least 150% of the calculated values (to be on the safe side).

Note that the power required from the 9V battery is $9\text{V} \times 66\text{mA} = 0.6\text{W}$, while the load may be consuming only $5.3\text{V} \times 10\text{mA} = 0.05\text{W}$, so the efficiency of our simple regulator is a measly $0.05/0.6 \approx 8\%$ (ouch!). This is typical for a Zener voltage regulator, which is called a *shunt regulator*: the current drawn from the source is always greater than the maximum required load current, even when the load is drawing little or no current most of the time.

> Zener diode voltage regulator circuits are really only practical when both the variation in the required load current (i_{load}) and the expected variation in the source voltage (v_{in}) are small.

For this particular problem a more efficient solution would be to use a *series regulator* (the standard type implemented by special-purpose voltage regulator ICs), or, even better, a *DC-*

DC converter, which is a type of *switching regulator* that can convert 9V power to 5V power with efficiencies exceeding 80%.

LEDs

A *light emitting diode* (LED) is a PN junction diode made from a semiconductor material with a relatively large energy gap. Consequently, when an electron and hole recombine the energy released may be carried away by a visible light (or near infrared) photon. Otherwise, LEDs behave much the same as any other semiconductor diode. Because of its larger gap voltage, the forward-bias voltage drop for an LED is significantly higher than for a silicon diode; typical forward voltages for various LEDs are:

Table 3-1

Typical LED Forward Voltage Drop (10mA forward current)

IR (950nm)	Red (630nm)	Yellow (590nm)	Green (565nm)	Blue (470nm)
1.2V	**1.8V**	**2.0V**	**2.2V**	**3.4V**

White LEDs are usually constructed using a blue emitter with a fluorescent coating, so their forward voltage drop is the same as for blue LEDs.

Caution

LED reverse breakdown voltage is typically only 5V! Be careful to avoid using an LED in a circuit which could cause it to suffer reverse breakdown.

Since the LED is a PN junction diode, *its forward current must be limited by the external circuit*; this is most often accomplished using a resistor in series with the diode.

The intensity of the light output of an LED is very nearly proportional to its forward-bias current; 10mA is usually more than sufficient to provide a nice, bright indicator light. Be careful to avoid excessive forward currents or the LED will quickly fail; if a very bright source is needed, specially designed and cooled LEDs are available (at a price). LED indicators are often controlled by digital logic circuitry, and are turned off or on depending on the logic circuit state; we'll discuss these sorts of circuits in a later experiment. In this section we consider some designs involving our analog circuits.

Often you will want an LED *pilot light* to illuminate whenever the power supply to a circuit is activated (like the LEDs on the circuit trainer breadboard). Some circuits to do this job are shown in Figure 3-19. For circuits (a) and (b) from that figure, the minimum voltage required for any significant illumination is the LED's turn-on voltage (Table 3-1). The resistor value is chosen so that the desired current will flow through the diode to give the desired intensity.

For example, illuminating a green LED with 10mA from a 12V source would require a resistor value of $R = (12\text{V} - 2.2\text{V})/10\text{mA} \approx 1\text{k}\Omega$; a 5V source would require $R \approx 270\Omega$.

The circuit in (c) is interesting because it uses a Zener diode to set the minimum voltage required for the LED to illuminate. This would be useful if you don't want the pilot light to illuminate if the power supply voltage is less than some threshold value. For example, assume you are using a 9V battery for your power supply, but the circuit needs at least 7V to operate properly. If the Zener breakdown voltage is 5.1V and you use a red LED (turn-on 1.8V), then at least $5.1\text{V} + 1.8\text{V} = 6.9\text{V}$ would be needed to illuminate the LED. With a current of 10mA at 9V, $R = (9\text{V} - 6.9\text{V})/10\text{mA} \approx 200\Omega$.

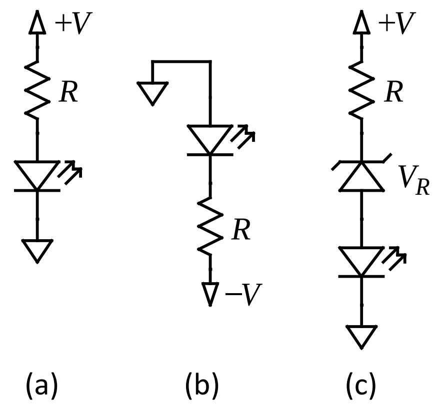


Figure 3-19: LED pilot light circuits. (a) positive supply voltage; (b) negative supply voltage; (c) positive voltage, but with a minimum voltage threshold for LED illumination using a Zener diode. See the text for details.

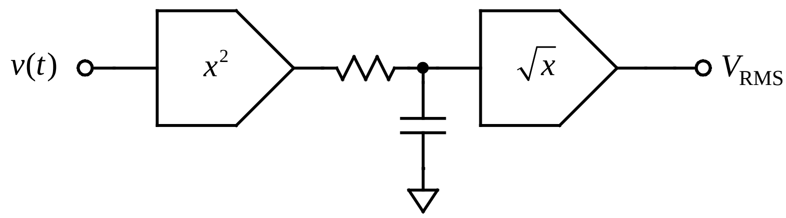
True RMS measurement using analog multipliers

The usual way to determine the power transmitted by an arbitrary, time-varying signal is to measure the average of its squared amplitude (where the average is over a time interval usually much longer than the period of the lowest frequency AC component in the signal). Consequently, the root-mean-square (*RMS*) amplitude of a waveform is typically used to characterize its magnitude:

3.5 $$V_{\text{RMS}} \equiv \sqrt{\left(\overline{v(t)^2} \right)}$$

For a DC voltage, V_{RMS} is just the DC voltage value; for a sinusoid, it is $1/\sqrt{2}$ of the phasor magnitude, or about 35% of its peak-peak voltage. For a complicated signal composed of a DC and several AC components, V_{RMS} is the Pythagorean sum (square root of the sum of the squares) of the individual component *RMS* values.

Since the mean of a time-varying signal is just its DC component (see equation 2.2, page 2-4 of Experiment 2), we can extract the mean of a signal using a low-pass filter with a cutoff frequency well below the lowest frequency of any AC signal component (the effective averaging time, for example, of a simple *RC* low-pass filter is approximately equal to its *RC* time constant). An analog circuit to output the RMS value of a time-varying input signal could then be constructed thusly:



The scale factors of the multipliers used for the square and square root circuits don't matter as long as they are the same; Figure 3-12 shows a square root circuit using the MPY634.

If you experiment with this circuit, use a variety of input waveforms with various frequencies and RMS amplitudes; compare the circuit's DC output voltage to the signal generator's RMS amplitude setting (make sure that the signal generator output setup is such that it reports the amplitude assuming a "High-Z" load). Estimate the circuit's accuracy and its high frequency response limit.

The PN junction diode

Insulators, conductors, and semiconductors

The electrical conductivities of solid materials for the most part fall into one of two classes: conductors and insulators (metals make up most of the conductors, and nonmetals are usually insulators). Although all materials are very nearly electrically neutral (equal numbers of protons and electrons, so that they carry no net charge), the nature of the chemical bonds which bind the atoms or molecules of a solid to one another determines its class of electrical conductivity.

An atom's *valence electrons* — the outer, most weakly bound electrons — are the ones which participate in chemical bonding. The atomic nucleus along with the much more strongly bound inner electrons comprise a positively-charged *ion core* which remains intact and is surrounded by the interacting valence electrons. The chemical bonding process causes these many ion cores in a solid to arrange themselves in a mostly regular, crystalline structure. This regular, periodic array of positively-charged cores creates a similarly regular, periodic electrostatic field within which the myriad valence electrons move.

The quantum-mechanical nature of these microscopic, negatively-charged particles (the valence electrons) as they evolve in the periodic electrostatic potential of the ion cores requires that they each occupy a state of motion (and total energy) in one of several distinct *energy bands*, analogous to the quantized energy states an electron may occupy in a single atom or molecule. The width of a typical energy band is on the order of a few to several electron volts (same order of magnitude as the binding energy of a valence electron in one of the atoms), and adjacent energy bands are often separated by a similar energy, although they may also overlap. Each band has enough distinct quantum states to contain twice the number of electrons as there are molecules in the macroscopic solid crystal (i.e. $\sim 10^{23}$).

Room temperature ($\approx 290\,\mathrm{K}$) corresponds to random, thermal particle energies of $\sim 1/40\,\mathrm{eV}$ (electron volt), much smaller than the width of an energy band but much larger than the energy spacing between the individual states in a band. Because electrons are subject to *Pauli Exclusion* (each electron must be in a unique, distinct quantum state), the valence electrons of all the various atoms in a solid fill the available states starting with the lowest available energy. Because room temperature corresponds to a fairly small energy, the energy of the topmost filled states is fairly well-defined and is called the electrons' *Fermi energy*. Random thermal jostling can only affect the states of individual electrons with energies near the Fermi energy, because those with much lower energies are surrounded by quantum states already occupied by other electrons, so they're stuck in their current states.

Now, one of two situations can occur for our valence electrons in a solid:

(1) The number of electrons is such that they exactly fill all the available states in some number of energy bands, and higher energy bands are completely empty.

(2) One energy band (or possibly more, if some bands overlap) is only partially filled and has many unoccupied states still available; all other bands are either completely filled or completely empty.

Electrons occupying a completely filled energy band do not participate in electrical conduction. The reason for this is that such a band corresponds to all physically possible states of individual electron motion in all directions consistent with the energies of the electrons in that band. Applying an external electric field doesn't change this situation unless the field is so intense that it can cause electrons to transition to another (partially filled or empty) energy band. Thus, no new net motion of electrons can be induced by the presence of the field, so the electrical conductivity contributed by a completely full (or, of course, completely empty) energy band is zero.

This last result implies that solids with situation (1) above are *insulators* (or maybe semiconductors). Since each band has twice the number of states as there are molecules in the crystal, insulating materials most often arise when there is an even number of valence electrons participating in the chemical bonding forming the solid. Situation (2), on the other hand, allows electrical conduction to proceed using the electrons in the partially-filled energy band. Electrons near the Fermi energy in the band have a wide selection of nearby empty states, so an applied electric field can accelerate them, and their resulting motions can carry a net flow of charge (electric current) through the solid. These materials are *conductors*, and partially-filled energy bands are characteristic of the so-called *metallic bond*.

Semiconductors have valence electrons whose situation falls into category (1): bands containing electrons are completely filled, at least at cold temperatures. What makes them different from insulators, however, is that the bottom of the nearest empty energy band (called the *conduction band*) is only about an eV or so away from the top of the highest-energy filled band (the *valence band*). Consequently, random thermal jostling of the ions in the lattice can impart enough energy to a few electrons with energies near the top of the valence band to excite them into levels near the bottom of the conduction band. In this case both the valence band and the conduction band become *partially* occupied (although just barely), and the material becomes a poor conductor (poor because only a tiny fraction of the valence electrons get bumped up into the conduction band). The higher the temperature, the greater the number of valence electrons thermally excited into the conduction band — the number goes as:

3.6
$$n_i \propto T^{3/2} e^{-E_g/(2k_B T)}$$

where E_g is the magnitude of the energy gap between the valence and conduction bands and k_B is Boltzmann's constant. In the case of silicon, this amounts to $\sim 10^9$ electrons per cm^3 at room temperature (compare with copper's 0.8×10^{23} per cm^3).

The archetypal semiconductors are the elements silicon ($E_g = 1.12\,\text{eV}$), and germanium ($E_g = 0.67\,\text{eV}$), each of which form a diamond lattice with four covalent bonds per atom. Carbon in its diamond form ($E_g = 5.5\,\text{eV}$) is beginning to find applications in solid-state devices, but its large energy gap makes it more properly classified as an insulator.[2] Several compounds and alloys form commercially important semiconductors, including GaAs, InP, GaAsP, and InGaN.

Electrons and holes; impurities and doping

The diagram at right illustrates the distribution of electrons between the top of the valence band and the bottom of the conduction band for a semiconductor at a fairly high temperature (so that there have been a considerable number of electrons excited into the conduction band). The density of quantum states grows as $\sqrt{\Delta E}$ as you move away from the band edges, as shown by the right-hand curves in the figure. The Boltzmann factor $\text{Exp}(-\Delta E/k_B T)$ gives the relative probability that any one state is occupied in the conduction band or unoccupied in the valence band. Because the number densities of the conduction electrons and the holes in their respective bands are low (much smaller than the crystal's atomic number density), the charge carriers will distribute themselves in accordance with the classical Maxwell distribution, so the overall occupation densities go as the left-hand curves in the figure.

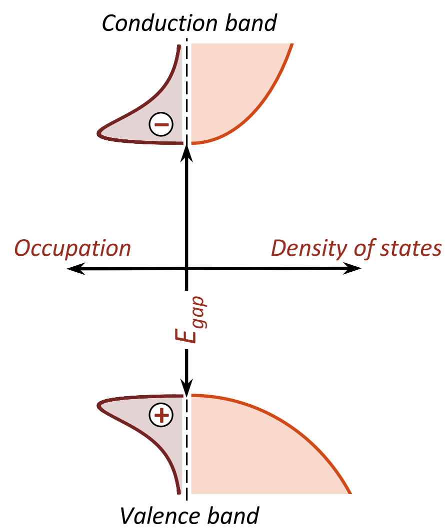


Figure 3-20: Densities of states and occupations by electrons (conduction band) and holes (valence band) for a pure semiconductor (only intrinsic charge carriers); energy increases in the vertical direction. The kinetic energy distribution of the charge carriers in each band is classical.

The dynamics of the relatively small number of electrons in the conduction band is very well approximated by treating them as classical particles (with a negative charge of $-q_e$, of course), but their *effective mass* is determined by the shape of the density of states curve near the bottom of the conduction band: the sharper the curve near the minimum, the lighter the effective mass. In the valence band, only a small fraction of the states near its top are unoccupied. Interestingly, the dynamics of the remaining

[2] Carbon in the form of graphite has layers of honeycomb arrangements of atoms (a single layer forms a 2-d structure called *graphene*). Graphite is a *semi-metal*: it has partially-filled bands like a conductor, but these bands are very nearly empty. Consequently the conduction electron density in graphite is much smaller than that of a typical conductor, 10^{18} vs. 10^{23} electrons per cm^3.

electrons near the top of the valence band are such that they have a *negative effective mass*, since the density of states *decreases* with increasing energy near the top of the band.

The consequence of this unusual electron behavior near the top of the valence band is that the unoccupied quantum states evolve as though they were *positively charged particles* $(+q_e)$ *with a positive effective mass and with energies increasing as they move further down from the band top*! These "positive charge carriers" near the top of the valence band are called *holes*. Thus, when an electron is excited from the valence band to the conduction band, *two* charge carriers are created: the electron $(-q_e)$ and the hole $(+q_e)$ it left behind. Since the energy an electron must gain to cross the energy gap between valence and conduction bands is at least E_g, but two "particles" were created by this transition, *the required energy per particle* is $E_g/2$ — this explains the 2 in the Boltzmann factor in equation 3.6.

Thus a pure semiconductor has a conductivity which is a very strong function of temperature, rising rapidly as temperature increases (equation 3.6). This effect is used to make a *thermistor*: a resistor with a large, *negative temperature coefficient* (decreasing resistance as temperature rises) which acts as a very sensitive, fast-acting temperature sensor for the range of about −100°C to +150°C.

Semiconductor materials are custom-made to be much more flexible and useful through the process of *doping*: introducing various amounts of impurity atoms into the semiconductor crystal which have a different valence than the semiconductor. For example, mixing a small amount of phosphorous (valence 5) into a silicon crystal will introduce atoms each with an extra valence electron left over after it forms bonds with surrounding silicon atoms. What would be the consequences of these extra electrons to the physics of the material? It turns out that the energy of this extra valence electron is very close to the energy of the bottom of the conduction band (in the case of P in Si, the energy is only 0.044eV below the conduction band). If there are relatively few of these *donor* impurity atoms, then it is very likely that such electrons will eventually be thermally excited into the conduction band: once there they move away from the impurity atoms and are unlikely to recombine with them.

So even if the ambient temperature is cool enough that almost no electrons would be excited from the valence band to the conduction band, electrons from donor impurities will nearly all find their way into the conduction band, providing a largely temperature-independent cadre of negative charge carriers $(-q_e)$ along with the same number of fixed, positively-charged ions $(+q_e)$ distributed throughout the crystal lattice. Such a material is called an *N-type semiconductor*.

Similarly, introducing a valence 3 impurity atom (such as aluminum into silicon) will leave an unsatisfied bond because of the missing electron. Again, the energy required to promote a nearby valence electron into this spot is small compared to the semiconductor's energy gap (0.057eV for Al in Si). Thermal agitation will eventually do the trick, and the vacated valence state becomes a hole which quickly moves away from the impurity atom, trapping the promoted electron at the impurity site. Thus these *acceptor* impurity atoms become fixed, negatively-charged ions $(-q_e)$ in the lattice, whereas an equal number of holes form a temperature-independent group of positive charge carriers $(+q_e)$, creating a *P-type semiconductor*.

Adding dopants to a semiconductor can not only introduce charge carriers (called *extrinsic* charge carriers), but will also suppress the thermal creation of electron-hole pairs described by equation 3.6 (called *intrinsic* charge carriers). This is because the product of the number of conduction electrons (n_c) and the number of holes (p_v) is related to the number of intrinsic charge carriers thermally created in a pure (undoped) semiconductor (n_i) by the laws of statistical mechanics:

3.7
$$n_c \, p_v = n_i^2$$

For example, the addition of 1 part per million phosphorous to a silicon crystal would introduce 5×10^{16} extrinsic conduction electrons per cm^3; with $n_i \sim 10^9$ electrons per cm^3, we see that there will be only $p_v \sim 100$ holes per cm^3! These holes are called *minority carriers* in the N-type silicon under discussion; the conduction electrons are the *majority carriers*. Since for this example $n_i \ll n_c$, the temperature dependence of n_c will be quite small, so, from equation 3.7, $p_v \propto n_i^2$. Thus *the temperature dependence of the minority carriers is very large*: from equation 3.7,

3.8
$$\langle\text{minority carrier density}\rangle \propto T^3 e^{-E_g/(k_B T)}$$

The equilibrium PN junction

Now consider the case of a semiconductor crystal with inhomogeneous doping. As a concrete (but quite artificial) example, assume that we take a single P-type crystal and a single N-type crystal and then join them along a planar boundary so as to form a single crystal with an abrupt change in doping at this boundary. The result is a *PN junction* (Figure 3-21) at the interface between the two semiconductor types.

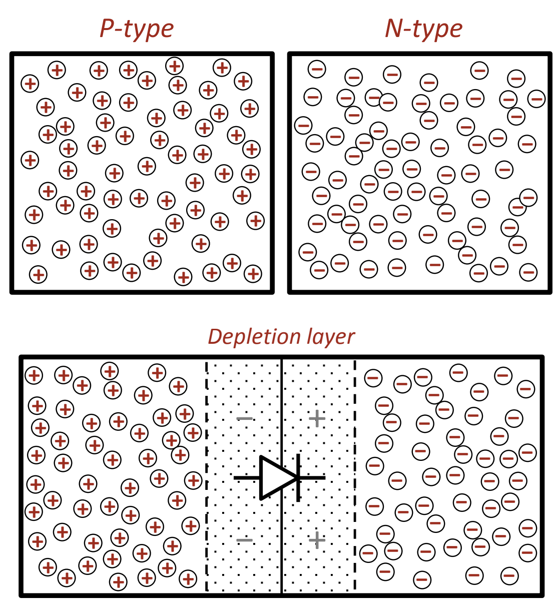


Figure 3-21: Formation of a PN junction diode and its depletion layer.

Far from the boundary the charge carrier densities must approach their homogeneous, thermal equilibrium values. Near the interface, on the other hand, the large gradients in the hole and conduction electron densities will drive diffusion of these charge carriers across the boundary, where they will eventually recombine with carriers of the opposite sign. The reduced majority carrier densities near the boundary induce a net charge density and resulting electric field near the interface because of the now unbalanced charge of the impurity ions in each semiconductor. This electric field will repel the majority charge carriers on either side of the boundary, and an equilibrium condition is reached preventing further net diffusion of carriers across the boundary (bottom illustration in Figure 3-21). The electric field near the boundary generates a potential difference between the P-type and N-type sides of the junction, with the N-type material at the higher potential. This *contact potential* of the PN junction is very close to the *gap voltage*: $E_g / q_e \equiv V_g$ (1.12V for silicon). The result, as we shall see, is the creation of the *PN junction diode*.

It turns out that the equilibrium situation will be attained only when a region near the PN interface is almost completely devoid of charge carriers: the so-called *depletion layer*, as shown in Figure 3-21. The charge density in this region is then given by the number densities of the impurity ions on each side of the boundary, which are nearly equal to the corresponding majority carrier number densities far from the boundary. Since the potential changes by $\sim V_g$ across the depletion layer, it is straightforward to calculate its equilibrium width, which will typically be in the range of $10^2 - 10^4$ Å (about 2000 Å for silicon with a part per million doping), and the magnitude of the electric field at the interface is in the range of $10^5 - 10^7$ V/m.

The PN junction I-V characteristic curve

The equilibrium configuration (Figure 3-21) is maintained when the rate that the holes diffuse into the depletion layer from the P-type side (against the contact potential gradient) matches the small rate of hole diffusion from the N-type side (where holes are the minority carriers), so that the net flow of holes across the junction is 0; a similar condition holds for the conduction electron diffusion at the junction.

Holes entering the depletion layer from the N-type side are not impeded by the presence of the contact potential — on the contrary, the electric field in the depletion layer will accelerate them through it to the P-type side. This implies that the rate of the minority hole diffusion will simply be proportional to the hole density on the N-type side, which is given in equation 3.8 to be proportional to $e^{-q_e V_g / (k_B T)}$, and similarly for the minority electron diffusion from the P-type side. The majority carriers must cross the barrier imposed by the junction potential (V_j), so only those carriers with kinetic energies larger than $q_e V_j$ can cross to the other side; the number of such energetic carriers will be proportional to the Boltzmann factor $e^{-q_e V_j / (k_B T)}$ (because their kinetic energy distributions are classical, as mentioned before). At equilibrium, these two rates match, and $V_{j0} \approx V_g$.

When an external bias voltage V is applied across the PN junction, this applied potential will reduce the junction potential to $V_j = V_{j0} - V$ ($V > 0$ is forward-biased). As a consequence, more majority charge carriers will have enough energy to diffuse through the depletion layer;

the minority diffusion rate from the other side is unaffected. Thus there will be a net current flow across the junction given by the difference in these two diffusion rates:

$$I \propto e^{-q_e (V_g - V)/k_B T} - e^{-q_e V_g / k_B T} = e^{-q_e V_g / k_B T} (e^{q_e V/k_B T} - 1)$$

This simple result is known as the *ideal diode equation*:

$$I = I_R (e^{q_e V/k_B T} - 1); \quad I_R = I_0 e^{-q_e V_g / k_B T}$$

V is the applied bias voltage (+ for forward-bias), V_g is the semiconductor's gap voltage, I_0 is some constant, and I_R is the diode's reverse leakage current. Thus, the ideal diode's forward current rises exponentially with forward bias voltage (for voltages of more than a few tens of millivolts), and has some small, temperature-dependent leakage current when reverse-biased.

The above equation is not quite right, because its derivation ignores an effect which is especially important for the behavior of a silicon diode: generation and recombination of charge carrier pairs in the depletion layer. The assumption in the argument leading up to the diode equation was that the only charge carriers present in the depletion layer entered it through diffusion from the regions outside the layer, and that all of these carriers pass through the depletion layer. Actually, thermal excitation of electron-hole pairs will occur in the depletion layer, just as it would in a pure semiconductor; similarly, recombination of electrons and holes may also occur among those which diffuse into the depletion layer, so the number of charge carriers entering the depletion layer is larger than the number which escape, especially for small forward bias voltages in relatively large V_g diodes such a silicon.

The depletion layer generation and recombination processes depend exponentially on temperature, but the exponent goes as $q_e/2k_B T$ rather than as $q_e/k_B T$. The combination of this process with the ideal diode process leads to a "slight" modification of the ideal diode equation:

Diode equation

3.9 $$\boxed{I = I_R (e^{q_e V/\eta k_B T} - 1); \quad I_R = I_0 e^{-q_e V_g / \eta k_B T}}$$

The coefficient η depends on the importance of the depletion layer recombination process; it is a weak function of I and T and ranges between 1 and 2. For the small-signal silicon diodes you will use, $\eta \approx 1.9$ and $I_R \approx 5\,\text{nA}$; $q_e/\eta k_B T \approx 20\,\text{volt}^{-1}$ at 20°C. The exponential dependence of I on forward-bias voltage V (for $I \gg I_R$) is the basis for the exponential and logarithmic amplifiers presented earlier; unfortunately, this relationship is strongly temperature-dependent. When the junction is reverse-biased, the depletion layer generation rate depends on the size of the depletion layer, which grows as $\sqrt{1+V/V_g}$ (V is the reverse-bias voltage), so the reverse current does not "saturate" at the I_R value given by equation 3.9, but continues to grow slowly with increasing reverse-bias voltage as long as it remains well below the diode's breakdown voltage. Cartoons of these diffusion processes are shown in

Figure 3-22; plots of the 1N4148 silicon small-signal diode I-V characteristic curves for both forward and reverse bias and at two temperatures are provided in Figure 3-23.

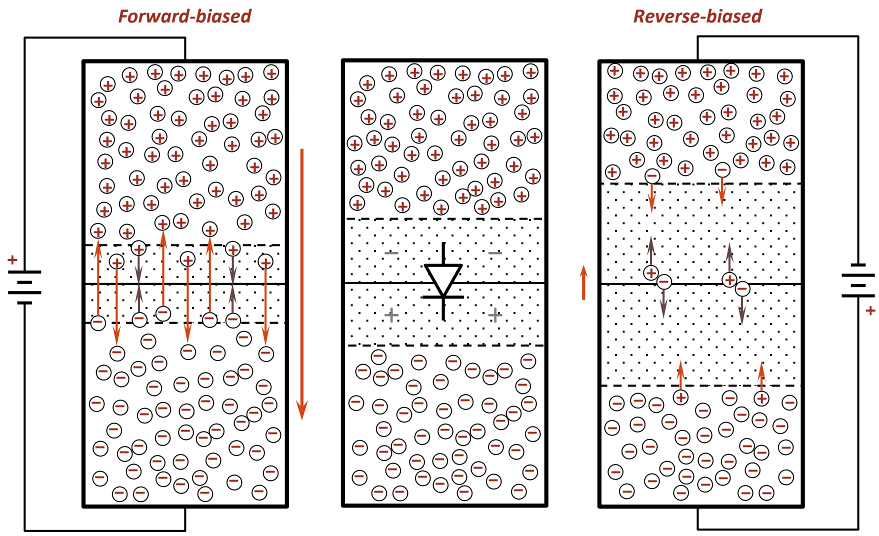


Figure 3-22: Depletion layer width and charge carrier diffusion of a PN junction as affected by applied bias voltage. Forward-bias (left) reduces the height of the potential barrier to majority carrier diffusion and decreases the depletion layer width, so many majority carriers can diffuse into and through the depletion layer; minority carrier diffusion is largely unaffected. Carriers that cross the depletion layer and recombine with majority carriers on the opposite side are indicated by the orange arrows; those that recombine inside the depletion layer are shown with gray arrows. Reverse-bias current is completely dominated by minority carrier diffusion: those that enter the depletion layer from the bulk semiconductor (orange arrows) and those pairs that are thermally generated within the depletion layer (gray arrows).

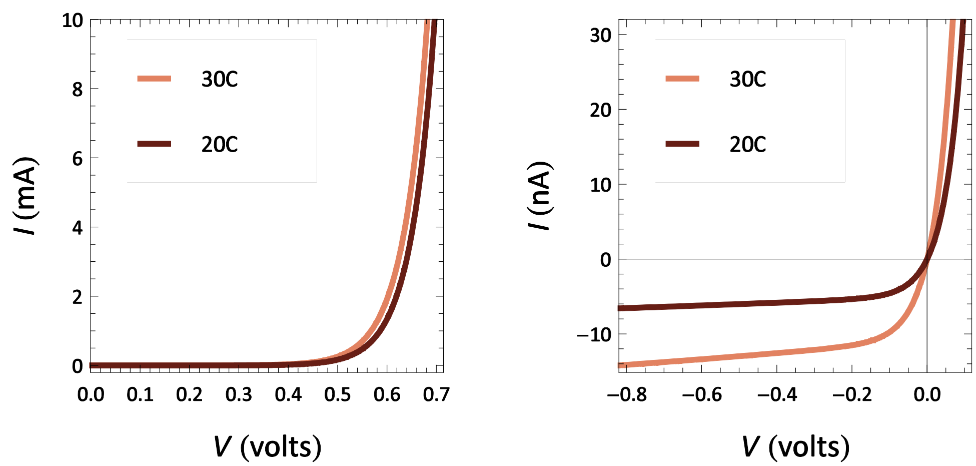


Figure 3-23: Forward-bias (left) and reverse-bias (right) I-V characteristic curves for the 1N4148 silicon diode. The forward-bias curves are exponential (equation 3.9), but they appear as though the diode suddenly “turns on” at a forward voltage of 0.6–0.7V; this voltage decreases slowly with rising diode temperature, as shown. The reverse-bias current is a much more sensitive function of temperature, however, approximately doubling for the same 10°C temperature increase. Note the different vertical scales for the two plots (a factor of 10^6).

Zener and avalanche breakdown

As the reverse-bias voltage on a PN junction is increased, the intensity of the electric field in the depletion layer rises; it is particularly intense at the interface between the P- and N-type areas. Minority carriers entering the depletion layer are accelerated by the field; when their kinetic energies reach a few eV or so, collisions with atoms in the lattice may knock valence electrons out of them, creating additional electron-hole pairs. These newly-created charge carriers are also accelerated by the field and can create even more carriers as they collide with lattice atoms.

At sufficiently high reverse voltages this collision-induced ionization process may lead to an *avalanche* of additional charge carriers, and the reverse current will grow exponentially with increasing voltage beyond some reverse-bias threshold. This is the *avalanche breakdown* process, and the reverse-bias voltage threshold for its action is the diode's *reverse breakdown voltage*. The electric field intensity for any particular applied reverse-bias voltage depends on the impurity concentrations and the abruptness with which these concentrations change near the P-type and N-type interface, so a target reverse breakdown voltage may be engineered into a particular diode type.

Another effect of a very intense electric field in the depletion layer is the large electric polarization of the atoms in the lattice it induces: at field strengths $\gtrsim 10^6\ \mathrm{V/m}$ the potential difference across a distance of about 100Å can exceed the semiconductor's gap voltage. In this case a valence electron may quantum mechanically *tunnel* across this distance into the conduction band, creating an electron-hole pair; because of this tunneling process the electric field required to ionize a lattice atom is much smaller than it would need to be to ionize a single, independent atom ($\sim 10^{10}$–$10^{11}\ \mathrm{V/m}$); this effect was first theorized by the American physicist Clarence Zener in 1934. The tunneling rate grows exponentially as the required tunneling distance (inversely proportional to electric field strength) decreases, again leading to a large increase in reverse current (breakdown) as applied reverse-bias voltage exceeds the tunneling threshold. The target reverse breakdown voltage may be engineered by adjusting a diode's impurity concentration and doping profile. See Figure 3-17 on page 3-20 for a typical reverse breakdown I-V characteristic.

Diodes with reverse-breakdown voltages exceeding 6V or so are dominated by the avalanche breakdown process; those below 5V are predominantly subject to Zener breakdown. Regardless of breakdown voltage, those diodes designed to be used as voltage regulators with precisely-tailored reverse breakdown voltages are collectively called *Zener diodes*; those with high current-handling capacity and extremely fast response to voltages exceeding their breakdown threshold are usually called *avalanche diodes* and are primarily used for *transient voltage suppression* (TVS) and overvoltage protection.