# Electronics for Experimenters 

Jeremiah Mans<br>University of Minnesota

October 12, 2011
(c) Copyright by Jeremiah Mans, 2011. All rights reserved.

## Contents

1 Introduction ..... 1
2 Passive Components ..... 3
2.1 Resistors ..... 3
2.1.1 Ohm's Law ..... 3
2.1.2 Resistor Networks and Kirchhoff's Rules ..... 4
2.2 Thevenin Equivalence and Characteristic Impedances ..... 6
2.2.1 Thevenin's Theorem ..... 6
2.2.2 Input and Output Impedance ..... 9
2.3 Capacitors and Inductors ..... 10
2.3.1 Capacitors ..... 10
2.3.2 Inductors ..... 11
2.4 RC and RL Transients ..... 11
2.5 RC Filters ..... 13
2.5.1 The High-Pass Filter ..... 14
2.5.2 The Low-Pass Filter ..... 16
2.5.3 General Characteristics of Simple Filters and Amplifiers ..... 17
2.5.4 RLC Circuits ..... 18
2.6 Real Components ..... 21
2.6.1 Real Resistors ..... 21
2.6.2 Real Capacitors and Inductors ..... 21
2.7 Diodes ..... 22
2.7.1 Diode Circuits ..... 23
2.7.2 Realistic Diode Characteristics ..... 24
3 Operational Amplifiers ..... 29
3.1 Amplification Principles ..... 29
3.2 Ideal Operational Amplifiers ..... 30
3.2.1 The Inverting Amplifier ..... 31
3.2.2 The Non-inverting Amplifier ..... 33
3.2.3 The Virtual Short Circuit ..... 35
3.3 Non-Ideal Op Amps ..... 36
3.3.1 Offset Voltage ..... 36
3.3.2 Bias Current ..... 37
3.3.3 Common-Mode Amplification ..... 39
3.3.4 Frequency Response ..... 39
3.3.5 Slew Rate ..... 41
3.4 Op-Amp Applications ..... 42
3.4.1 Summing and Differencing Amplifiers ..... 42
3.4.2 The Instrumentation Amplifier ..... 44
3.4.3 Current to Voltage and Voltage to Current ..... 45
3.4.4 Integrators and Differentiators ..... 48
3.4.5 Comparators ..... 50
4 Sensors ..... 55
4.1 Basic Sensor Principles ..... 55
4.1.1 Voltage and Current Sensors ..... 56
4.1.2 Resistive Sensors ..... 57
4.1.3 Capacitive and Inductive Sensors ..... 59
4.1.4 Semiconductor Device Physics ..... 60
4.2 Selected Sensors ..... 61
4.2.1 Temperature Sensors ..... 61
4.2.2 Photosensors ..... 66
4.3 Advanced Measurement Techniques ..... 67
4.3.1 The Lock-In Amplifier ..... 67
5 Transistors ..... 69
5.1 Basic Transistor Operation ..... 69
5.1.1 Bipolar Junction Transistors ..... 70
5.1.2 Field-Effect Transistors ..... 72
5.2 Use of Transistors ..... 73
5.2.1 Switching ..... 73
5.2.2 Linear Power Boosting ..... 76
6 Digital Electronics ..... 79
6.1 Combinatorial Logic ..... 79
6.1.1 Boolean Logic ..... 80
6.1.2 A Full Set of Gates ..... 82
6.1.3 Decoders and Multiplexers ..... 85
6.1.4 Memory: The Universal Gate ..... 87
6.2 Logic Standards ..... 89
6.2.1 Positive Voltage Logic Families ..... 89
6.2.2 Shared Lines ..... 91
6.2.3 Differential and Specialized Standards ..... 93
6.3 Sequential Logic ..... 94
6.3.1 Limits of Combinatorial Logic ..... 94
6.3.2 D Flip-Flop ..... 96
6.3.3 Counters ..... 97
6.4 State Machines ..... 99
7 Programmable Digital Logic ..... 105
7.1 Programmable Logic Devices ..... 105
7.1.1 PALs and CPLDs ..... 105
7.1.2 FPGAs ..... 108
7.2 Programmable Logic Design ..... 110
7.2.1 Basics of Verilog ..... 111
7.2.2 Sequential Logic ..... 112
7.2.3 Internal signals ..... 113
7.2.4 State Machine Example ..... 113
8 Digital and Analog Together ..... 115
8.1 Digital to Analog Conversion ..... 115
8.1.1 Summing DAC ..... 116
8.1.2 R-2R Ladder ..... 117
8.1.3 Pulse-Width Modulation ..... 118
8.1.4 Sigma-Delta Conversion ..... 119
8.2 Analog to Digital Conversion ..... 122
8.2.1 ADC Helper: the Sample-and-Hold ..... 122
8.2.2 Flash ADC ..... 123
8.2.3 Integration ADC ..... 123
8.2.4 Successive-Approximation ADC ..... 125
8.3 Time to Digital Conversion ..... 127
8.3.1 Pure Digital Techniques ..... 127
8.3.2 Analog-based Techniques ..... 128

## List of Figures

2.1 Schematic symbols for resistors ..... 4
2.2 Two resistors connected in series, forming a voltage divider. ..... 4
2.3 Two resistors connected in parallel, forming a current divider. ..... 6
2.4 Thevenin equivalent circuit ..... 8
2.5 Input and output impedances connected together. ..... 9
2.6 Schematic symbols for (a) unpolarized capacitors, (b) polarized capacitors, and (c) inductors. ..... 10
2.7 RC and RL circuits ..... 12
2.8 RC and RL Transients as a function of time. ..... 14
2.9 High-pass RC filter ..... 15
2.10 Low-pass RC filter ..... 16
2.11 Bode Plot of a high-pass RC filter. ..... 17
2.12 An RLC filter ..... 19
2.13 Bode plot for the notch filter ..... 20
2.14 Schematic symbol for a diode ..... 22
2.15 A simple diode rectifier circuit (a) and $V(t)$ behavior (b). ..... 23
2.16 A diode rectifier circuit including a filter capacitor (a) and the resulting $V(t)$ behavior (b) ..... 24
2.17 Characteristic diode current versus voltage behavior (IV curve). The mag- nitude of $I_{r l}$ is generally $\mathrm{O}(10 \mathrm{nA})$ or smaller. ..... 24
2.18 A simple Zener-diode-based voltage regulator. ..... 25
3.1 Schematic symbol for an operational amplifier showing the power supply terminals. ..... 30
3.2 The inverting amplifier. ..... 31
3.3 The non-inverting amplifier. ..... 33
3.4 The op-amp buffer or unity-gain amplifier. ..... 34
3.5 A meter amplifier. ..... 35
3.6 A model of offset voltage. ..... 37
3.7 Cancellation of bias current effects in the inverting amplifier. ..... 38
3.8 Waveform distortions induced by slew limits. In both figures, the dotted lines show the expected response from the circuit in the absence of slew limits while solid curves show the actual response. ..... 41
3.9 The summing or averaging amplifier. ..... 42
3.10 The differencing amplifier. ..... 43
3.11 The instrumentation amplifier. ..... 45
3.12 A transimpedance or current-to-voltage amplifier. ..... 46
3.13 A voltage-to-current amplifier. ..... 47
3.14 An op-amp based integrator circuit with reset ..... 48
3.15 Circuit designs for a simple (a) and realistic (b) differentiator using an op-amp. ..... 49
3.16 An op-amp-based comparator. ..... 50
3.17 An open-collector output comparator allowing connection to any digital logic family. ..... 51
4.1 Reference-line technique for subtracting common-mode noise using an in- strumentation amplifier. ..... 56
4.2 Direct constant-current (a) and constant-voltage (b) techniques for resistive- sensor measurements. ..... 57
4.3 Wheatstone bridge and instrumentation amplifier combination for precise measurement of small changes in resistance. ..... 58
4.4 Measurement of a changing capacitance using a high-pass filter technique. ..... 60
4.5 Band-diagram structure for a pure-semiconductor (a), doped semiconductor at $T=0 \mathrm{~K}(\mathrm{~b})$, doped semiconductor at finite temperature. ..... 61
4.6 A pn-junction ..... 62
4.7 Temperature measurement using iron/constantan $(\mathrm{CuNi})$ thermocouples with an ice-bath as a reference temperature. ..... 64
4.8 Example circuit showing the use of an AD590 semiconductor temperature sensor with an output scale of $1 \mathrm{mV}=1 \mathrm{~K}$. ..... 65
4.9 Light amplitude measurement with a photodiode and transimpedance am- plifier. ..... 67
5.1 Schematics and current flow labels for npn (left) and pnp (right) bipolar junction transistors. ..... 70
5.2 Common-emitter npn amplifier ..... 71
5.3 Schematic symbols for n-channel and p-channel MOSFET transistors. ..... 72
5.4 Switching a load using an n-channel MOSFET, with diode production for inductive loads ..... 74
5.5 Switching a load using an p-channel MOSFET, with diode production for inductive loads. ..... 75
5.6 Increasing the current (power) capacity of a non-inverting amplifier using an npn transistor. ..... 76
5.7 Increasing the current (power) capacity of a non-inverting amplifier using a push-pull stage. ..... 77
6.1 Implementation of the NOT (a), AND (b), and OR (c) functions using only NAND gates, demonstrating the universal nature of the NAND gate. ..... 83
6.2 Detail of the protection diodes included on modern digital logic inputs. ..... 90
6.3 Schematic representation of a push-pull or totem-pole digital logic output. ..... 91
6.4 Schematic representation of a tristate logic buffer where the $T$ input can be used to push the output to high-impedance. ..... 92
6.5 Schematic representation of an open collector output driver. The pull-up resistor is usually provided externally to the device. ..... 93
6.6 Example circuit for glitch-analysis. ..... 95
6.7 Timing diagram analyzing the behavior of the circuit in Figure 6.6. ..... 96
6.8 Design of a 3-bit ripple counter. ..... 98
6.9 Timing analysis of a 3-bit ripple counter. ..... 99
6.10 Implementation of 3-bit synchronous counter. ..... 100
7.1 Programmable AND-OR tree as used by Programmable Array Logic/Generic Array Logic devices. ..... 107
7.2 Structure of a Xilinx Coolrunner-II CPLD. The "PLA" is the PAL-like logic block while the "AIM" is the interconnection matrix between the logic blocks. Distribution of clocks and global resets is also shown. ..... 108
7.3 Individual logic cell of an FPGA ..... 109
7.4 Structure of a Spartan-IIE Input-Output Block (IOB) ..... 110
8.1 A digital-to-analog converter based on a summing amplifier configuration. ..... 117
8.2 A digital-to-analog converter based on an R-2R ladder summing amplifier configuration. ..... 117
8.3 Pulse-train output for a 4-bit PWM for a range of different input digital values. ..... 118
8.4 Accumulator sum and output pulse train for a 4-bit sigma-delta DAC for two different input digital values (4 and 14). ..... 122
8.5 Example of a sample-and-hold circuit. ..... 123
8.6 A two-bit flash-ADC. ..... 124
8.7 The core integrator and comparator for a dual-slope ADC. ..... 124
8.8 DAC output voltages during a four-bit successive approximation process. At each clock, a comparator determines if the input value is above or below the DAC value, determining the DAC value for the next step. ..... 126
8.9 Current-integrating analog TDC ..... 128

## Chapter 1

## Introduction

The development of the transistor, integrated circuits, and modern computers have had a major impact on our daily lives. Electronics surround us - the average student is generally carrying at least two highly complex electronic devices everywhere he or she goes.

Electronics have also had a significant impact on science, and much of modern science depends criticially on electronics techniques. For an experimenter, electronics are a tool to perform measurements and record results. Electronic systems can perform automated measurements, control the environment in which experiments take place, and even help analyze the resulting data.

The discipline of electrical and computer engineering is entirely devoted to the understanding and development of electronics. An electrical engineer is expected to devote years of study to cover the full range of electronic phenomena and tools. Most experimenters in other disciplines do not have the time to devote to a complete study, but rather need a working knowledge tuned to the needs of scientific measurement. The goal of this text is to support the education of experimenters in the most important techniques and tools in digital and analog electronics, with a focus on making physical measurements using electronic techniques.

## Chapter 2

## Passive Components

The most numerous components in modern electronics designs are not the fancy integrated circuits (chips) but rather the humble "passive" components such as resistors, capacitors, and inductors. These components provide the integrated circuits with the appropriate environment in which to work and tailor their behavior for the specific task at hand. In fact, knowledge of basic concepts of resistors, capacitors, and inductors is critical for understanding how these circuits operate. These topics are generally included in introductory physics courses, but we will review them here as they will be needed for understanding amplifier circuits and other later topics. In fact, we will see one particular circuit (the voltage divider) so often that one might describe any course in analog electronics as a course in "applied voltage dividers."

### 2.1 Resistors

### 2.1.1 Ohm's Law

The simplest component to discuss is the purely dissipative element: the resistor. A resistor is a two-terminal device which exhibits a linear relationship between the voltage applied across the device and the current which flows in this situation. The constant of proportionality is called the resistance and the linear equation is given the title Ohm's Law:

$$
V=I R
$$

Thus, a high-value resistor will allow a smaller current to flow at a given voltage than a low-value resistor. It is important to remember that the voltage in Ohm's Law is the voltage difference across the terminals, not an absolute voltage with respect to a global ground. For this reason, it can be helpful to write Ohm's Law more explicitly as

$$
\Delta V=I_{R} R
$$

where $I_{R}$ is the current through the resistor and $\Delta V$ is the voltage drop across it.

(a)

(b)

Figure 2.1: Schematic symbols for resistors

The implications of a larger or smaller resistor for power dissipation can be determined by combining Ohm's Law with the electrical power equation: $P=I V$. This yields two separate equations, depending on whether the circuit is at constant voltage or constant current.

$$
\begin{array}{ll}
P=I V=\frac{V^{2}}{R} & (\text { constant voltage }) \\
P=I V=I^{2} R & (\text { constantcurrent })
\end{array}
$$

Therefore, a large-value resistor dissipates less power than a small-value resistor in a constantvoltage circuit while it dissipates more power in a constant-current application.

In circuit diagrams, a resistor is shown as either a wavy line (Figure 2.1a) or a rectangle (Figure 2.1b). The wavy line is more common in the United States and Canada, while the rectangular symbol is the European standard.

### 2.1.2 Resistor Networks and Kirchhoff's Rules



Figure 2.2: Two resistors connected in series, forming a voltage divider.
Resistors are often used in combinations or networks in which multiple resistors are
connected together. The simplest such network is the series resistor combination seen in Figure 2.2. In this circuit, two resistors ( $R_{1}$ and $R_{2}$ ) are each connected at a single terminal with a voltage source across the pair. The connection between the two resistors is an interesting node in the circuit, which will have its own voltage $V^{\prime}$.

The behavior of resistor networks can be difficult to calculate, but the process is simplified by the use of Kirchhoff's Rules. These rules can be easily derived from Maxwell's equations, but are cast in a form which is simpler to apply to circuit problems.

- Kirchhoff's Current Rule - For any node in the circuit, the total sum of the currents must be zero. That is, the currents into the node must be balanced by currents out of the node. This rule guarantees that there is no net buildup of charge in the circuit with time.
- Kirchhoff's Voltage Loop Rule - The directed sum of voltages around any closed loop must be zero. For this calculation, the direction of current flow must be defined. Then, one must follow the loop in question. When following the loop in the direction of current flow, resistors give a voltage drop and batteries (or similar voltage sources) give a voltage gain.

The current rule should be used to resolve ambiguities about junctions while the voltage rule can be used to resolve questions about loops within the circuit network.

We can easily apply Kirchhoff's rules to the series resistor case. In this case, there is no loop so the relevant rule is the current rule.

$$
\begin{aligned}
I_{1} & =I_{2}=I \\
\frac{V-V^{\prime}}{R_{1}} & =\frac{V^{\prime}}{R_{2}} \\
\frac{V}{R_{1}} & =\frac{V^{\prime}}{R_{1}}+\frac{V^{\prime}}{R_{2}}=V^{\prime} \frac{R_{1}+R_{2}}{R_{1} R_{2}} \\
V^{\prime} & =V \frac{R_{2}}{R_{1}+R_{2}} \\
I & =\frac{V}{R_{1}+R_{2}}
\end{aligned}
$$

From this analysis, we learn two things. First, the combination of two resistors in series behaves like a single resistor with a total resistance $R=R_{1}+R_{2}$. Second, the voltage at the internal node depends on the ratio of resistors and not their absolute values. If $R_{1} \gg R_{2}$, then the voltage will be close to ground while if $R_{2} \gg R_{1}$ it will be close to $V$. If $R_{1}=R_{2}$, then $V^{\prime}=V / 2$. This behavior explains why the series resistor combination is often called a voltage divider.

Kirchhoff's rules can also be applied to the other classic simple resistor network: resistors in parallel as in Figure 2.3. This case is perhaps even easier to analyze. The voltage drop across the two resistors must be the same by Kirchhoff's loop rule, so the current across each is simply determined by the applied voltage.

$$
I_{1}=\frac{V}{R_{1}}
$$



Figure 2.3: Two resistors connected in parallel, forming a current divider.

$$
I_{2}=\frac{V}{R_{2}}
$$

Kirchhoff's current rule tells us that the total current must simply be the sum of the currents through the resistors. This allows us to derive the effective resistance of the network.

$$
\begin{aligned}
I & =I_{1}+I_{2} \\
& =\frac{V}{R_{1}}+\frac{V}{R_{2}} \\
& =V \frac{R_{1}+R_{2}}{R_{1} R_{2}} \\
R_{e f f} & =\frac{R_{1} R_{2}}{R_{1}+R_{2}}
\end{aligned}
$$

While the series resistor configuration served as a voltage divider, the parallel configuration operates as a current divider. This can be easily seen by looking at the ratio of the currents through the two resistors.

$$
\frac{I_{1}}{I_{2}}=\frac{\frac{V}{R_{1}}}{\frac{V}{R_{2}}}=\frac{R_{2}}{R_{1}}
$$

As one would expect, the larger current will flow through the smaller resistance.

### 2.2 Thevenin Equivalence and Characteristic Impedances

### 2.2.1 Thevenin's Theorem

Now that we have reviewed the equivalent resistances for series and parallel resistances, we can introduce Thevenin's Theorem which states that any combination of resistors and voltage sources which results in two terminals can be represented as single voltage source with voltage $V_{e}$ in series with a single resistor $R_{e}$. The Thevenin circuit is shown in Figure 2.4.

The values of $V_{e}$ and $R_{e}$ for a given network can be determined in several ways. In the simplest, consider a test resistor $R_{t}$ placed across the terminals of circuit. We can then

## Special Topic: Measuring Small Resistances

Resistances are generally measured using Ohm's Law. For a simple multimeter, one contacts the two leads of the device with the probes of the multimeter. Internally, the multimeter provides a known voltage across the probes and measures the current which flows through the probes, as in (a) below. The resistance measurement is thus $R_{m}=\frac{V}{I}$.
However, the resistance of the probes themselves and the contacts between the probes and device under measurement must be considered as well. The probe and contact resistances appear in series with $R$ so we will lump them into a single effective resistance $\left(R_{p}\right)$ as seen in (b). The true Ohm's Law rule is then

$$
R_{m}=R+R_{p}=\frac{V}{I}
$$

If we assume $R=R_{m}$, we make a fractional error

$$
\frac{\delta R}{R}=\frac{R_{m}-R}{R}=\frac{R_{p}}{R}
$$

If $R \gg R_{p}$, we can generally neglect the effect of probe resistance. However, when the resistance is small (for example when trying to measure superconductivity transitions) the effect of the probe and contact resistances can be significant. If we knew $R_{p}$ well, we could subtract it. Unfortunately, $R_{p}$ is difficult to determine and often changes with time as a result of environmental factors.
The alternative solution is the four-point resistance measurement. In the four-point technique, separate leads are used to provide the current and to measure the voltage across the sample. The four-point technique is shown in (c). The key feature of the four-point technique is that very little current flows in the loop containing the voltmeter. Four-point voltage measurements frequently have resistances in the $G \Omega$ range, which means that the addition of the voltmeter in parallel to the device under measurement has a minimal effect on the effective resistance of the system. Four-point measurement techniques are widely used in low-temperature condensed matter physics measurements where the leads to a device are necessarily long and thin to reduce heat conduction into a cryogenic volume.



Figure 2.4: Thevenin equivalent circuit
determine $V_{e}$ by setting $R_{t} \rightarrow \infty$ and analyzing the resultant circuit. The "open-circuit" voltage observed across $R_{t}$ is the Thevenin equivalent voltage $V_{e}$.

We can then determine $R_{e}$ by taking $R_{t} \rightarrow 0$ and determining the "short-circuit" current $I_{s c}$ which flows through $R_{t}$ in this case. The Thevenin equivalent resistance can then be determined using Ohm's Law:

$$
R_{e}=\frac{V_{e}}{I_{s c}} .
$$

## Example

Let us calculate the Thevenin equivalent voltage and resistance of the circuit below:


Since there is a single voltage supply of 6 V , it is tempting to make the assignment that $V_{e}=6 \mathrm{~V}$ but that is incorrect. We can determine the proper value for $V_{e}$ by leaving the terminals open and calculating the voltage across the output. In this case, we have a voltage divider, so

$$
\begin{aligned}
V_{e} & =V \frac{R_{2}}{R_{1}+R_{2}} \\
& =6 \mathrm{~V} \cdot \frac{400 \Omega}{200 \Omega+400 \Omega} \\
& =4 \mathrm{~V}
\end{aligned}
$$

Now, we can short-circuit the output terminals and determine the current which flows. In this case, the full voltage drop will appear across the $200 \Omega$ resistor, so it is again tempting
to make the assignment $R_{e}=200 \Omega$. Again, this would be incorrect as we can see by following the procedure described above.

$$
\begin{aligned}
I_{s c} & =\frac{6 \mathrm{~V}}{200 \Omega}=30 \mathrm{~mA} \\
R_{e} & =\frac{V_{e}}{I_{s c}}=\frac{4 \mathrm{~V}}{30 \mathrm{~mA}} \\
& =133 \Omega
\end{aligned}
$$

$$
\text { Thus } V_{e}=4 \mathrm{~V} \text { and } R_{e}=133 \Omega \text {. }
$$

### 2.2.2 Input and Output Impedance

We can extend the idea of Thevenin equivalent circuits to a pair of critical concepts for electronics and experiment design: input impedance and output impedance. Impedance is a generalization of resistance which can be applied to capacitors and inductors as well as to resistors. Impedance is generally given the symbol $Z$. The concepts of input impedance and output impedance are useful for simply characterizing devices which may be quite complicated. The model taken is of a single-port device, with a voltage output or input referenced to a circuit ground.

Let us begin with output impedance. Every circuit which provides a signal can be modeled with a Thevenin equivalent circuit. The $V_{e}$ represents the signal while the $R_{e}$ indicates the ability of the circuit to provide current at the signal voltage and is termed the output impedance $Z_{o}$. The larger the value of $Z_{o}$, the weaker the drive - since $Z_{o}$ appears in series with any later circuit it behaves like $R_{1}$ in a voltage divider. When $R_{1}>R_{2}$, the output voltage will be much lower than the original signal voltage. An ideal signal source will have an output impedance close to zero.

Input impedance can also be modeled by a Thevenin equivalent circuit: one where $V_{e}=$ 0 . Thus, the device is modeled as a resistor to ground and $Z_{i}$ is the value of the resistor. An


Figure 2.5: Input and output impedances connected together.


Figure 2.6: Schematic symbols for (a) unpolarized capacitors, (b) polarized capacitors, and (c) inductors.
ideal input usually has a large input impedance - the matching of a low output impedance to a high input impedance maximizes voltage signals. However, input impedances may also be designed to provide proper termination of long cables - an abrupt change of impedance between the cable and the device input will cause reflections.

On an intuitive level, we can consider the paired output impedance and input impedance as behaving as a voltage divider as in Figure 2.5. The signal seen on the connecting line will be

$$
V_{o}=V_{i} \frac{Z_{i}}{Z_{o}+Z_{i}}=V_{i} \frac{1}{1+\frac{Z_{o}}{Z_{i}}}
$$

so to maximize the output signal, we must make $Z_{o}$ small or $Z_{i}$ large or a combination of the two.

### 2.3 Capacitors and Inductors

Capacitors and inductors are relatively uninteresting devices in pure DC-circuits, but they are quite important for the majority of circuits which have significant time-variation in input and output. The schematic symbols for these devices are given in Figure 2.6. Capacitors and inductors should be familiar from previous courses so we will proceed directly to write down the equivalent statements to Ohm's Law for capacitors and inductors.

### 2.3.1 Capacitors

For capacitors, the charge on a capacitor is proportional to the voltage across it. By applying a time derivative to both sides, we find that

$$
I=C \frac{d V}{d t}
$$

If we apply an alternating voltage to the capacitor of the form $V(t)=V_{0} e^{i \omega t}$, the resulting current will be

$$
I(t)=C V_{0} i \omega e^{i \omega t}=V(t) i C \omega
$$

Thus, by analogy to Ohm's law, we can define the impedance of the capacitor to be

$$
Z_{C}=\frac{-i}{C \omega}
$$

The complex portion of the impedance indicates the phase shift induced by the capacitor. The standard rules for parallel and series resistors can be directly extended to complex impedances and the final complex angle determines the phase shift of the combined system. We will investigate this effect more fully when we consider filters.

The limiting cases for a capacitor's impedance are useful to remember: for low frequencies the effective resistance goes to infinity, while for high frequencies the capacitor become a short circuit with no resistance at all.

### 2.3.2 Inductors

The voltage across an inductor, by contrast, depends on the change in the current.

$$
V=L \frac{d I}{d t}
$$

If we apply a time-dependent current $I(t)=I_{\circ} e^{i \omega t}$, we can determine the voltage as a function of time and the magnitude of the voltage.

$$
\begin{aligned}
V(t) & =L \frac{d I}{d t}=L I_{0} i \omega e^{i \omega t} \\
& =i L \omega I(t)
\end{aligned}
$$

Therefore the impedance of an inductor can be defined as

$$
Z_{L}=i L \omega .
$$

If we consider the limiting cases for the inductor, at low frequencies the impedance goes to zero while at high frequencies the impedance becomes infinite. This is the opposite frequency behavior to the capacitor. The zero impedance at zero frequency is an idealized inductor behavior, the behavior of realistic inductors is discussed further below.

### 2.4 RC and RL Transients

Realistic and interesting circuits are rarely built of only one type of component. Instead, the most interesting circuits contain both resistive components and reactive components such as capacitors and inductors.

Let us begin our analysis of compound circuits by studying the behavior of the two circuits in Figure 2.7 in response to a step change of voltage. In both circuits, a resistor is connected to voltage supply in series with either a capacitor (a) or an inductor (b). We will assume that the applied voltage was initially zero $(V(t ; t<0)=0)$ and at $t=0$ it was raised to a constant value $\left(V(t ; t \geq 0)=V_{0}\right)$. The initial voltage across the capacitor and current through the inductor are thus both zero.


Figure 2.7: RC and RL circuits

Let us analyze the capacitor situation first. The circuit itself behaves like a voltage divider, with the capacitor as one of the resistors. Thus we can follow the same analysis as before, using Kirchhoff's Current Sum rule.

$$
\begin{aligned}
I_{R} & =I_{C} \\
\frac{V_{\circ}-V_{c}}{R} & =C \frac{d V_{c}}{d t} \\
& =-C \frac{d\left(V_{\circ}-V_{c}\right)}{d t}
\end{aligned}
$$

We can easily solve this differential equation by making the substitution $Q \equiv V_{\circ}-V_{C}$.

$$
\begin{aligned}
\frac{Q}{R} & =-C \frac{d Q}{d t} \\
-\frac{d t}{R C} & =\frac{d Q}{Q} \\
-\frac{t}{R C}+A & =\ln Q \\
V_{\circ}-V_{C} & =B e^{-t / R C} \\
V_{c} & =V_{\circ}-B e^{-t / R C}
\end{aligned}
$$

We are left with a constant of integration which we can resolve by observing that that $V_{c}(0)=0$, so

$$
V_{c}(t)=V_{\circ}\left(1-e^{-t / R C}\right)
$$

Thus, the voltage across the capacitor will exponentially approach the input voltage as it charges through the resistor. The rate of approach is determined by a time constant $\tau=R C$, so a larger resistance or larger capacitance will increase the time constant and slow the voltage rise. In later chapters, we will use this behavior to create oscillators with frequencies determined by a simple RC circuit.

For the inductor case we will naturally obtain relationships for currents rather than voltages, but the mathematics is similar. We begin by using Kirchhoff's Voltage loop rule
to write

$$
\begin{aligned}
\Delta V_{R}+V_{L} & =V_{\circ} \\
I R+L \frac{d I}{d t} & =V_{\circ} \\
\frac{L}{R} \frac{d I}{d t} & =\frac{V_{\circ}}{R}-I
\end{aligned}
$$

Since $\frac{d\left(V_{o} / R\right)}{d t}=0$, we can make the assignment $Q \equiv \frac{V_{0}}{R}-I$ and obtain

$$
\begin{aligned}
-\frac{d Q}{d t} & =\frac{R}{L} Q \\
\frac{d Q}{Q} & =-\frac{R}{L} d t \\
\ln Q & =-\frac{R}{L} t+A \\
Q & =A e^{-\frac{R}{L} t} \\
I & =\frac{V_{\circ}}{R}-A e^{-\frac{R}{L} t}
\end{aligned}
$$

Again, we use the initial condition $I(0)=0$ to set the constant of integration, so that we learn

$$
I(t)=\frac{V_{\circ}}{R}\left(1-e^{-\frac{R}{L} t}\right)
$$

For comparison with the capacitor case, however, we should convert this into the voltage across the inductor.

$$
\begin{aligned}
I R+V_{L} & =V_{\circ} \\
V_{\circ}\left(1-e^{-\frac{R}{L} t}\right)+V_{L} & =V_{\circ} \\
V_{L} & =V_{\circ} e^{-\frac{R}{L} t}
\end{aligned}
$$

Thus, for a voltage transient a capacitor will initially have a large inrush current and a low voltage across it while an inductor will have large voltage across its terminals and a low current. As time proceeds, both cases will approach their long-term average behavior as a short circuit (inductor) or an open circuit (capacitor). The voltage behavior as a function of time for both circuits is shown in Figure 2.8.

### 2.5 RC Filters

Many interesting signals are not simple transients, but rather are periodic signals or can be expressed (via Fourier analysis) as a sum of periodic signals. Combinations of resistors and capacitors or inductors can be used to attenuate periodic signals in a way which depends on the wavelength or frequency of the signal, a process called filtering.


Figure 2.8: RC and RL Transients as a function of time.

The actual attenuation of a filter will depend on frequency and is often expressed using the decibel ( dB ). The decibel attenuation $A$ is defined in terms of power, which for a constant impedance system will be proportional to $V^{2}$.

$$
A=10 \log _{10}\left(\frac{P_{\text {out }}}{P_{\text {in }}}\right)=10 \log _{10}\left(\frac{Z V_{\text {out }}^{2}}{Z V_{\text {in }}^{2}}\right)=20 \log _{10}\left(\frac{V_{\text {out }}}{V_{\text {in }}}\right)
$$

An attenuation of -3 dB thus corresponds to a power ratio of $50 \%$ or a voltage ratio of $\approx 71 \%$.

### 2.5.1 The High-Pass Filter

Consider the circuit in Figure 2.9 which is another voltage divider, this time with the capacitor standing in for $R_{1}$ instead of $R_{2}$. If we apply a periodic signal $V(t)$ with a characteristic angular frequency $\omega$, we would expect to observe a periodic $V^{\prime}(t)$ at the output node.

$$
\begin{aligned}
I_{C} & =I_{R} \\
C \frac{d\left(V(t)-V^{\prime}(t)\right)}{d t} & =\frac{V^{\prime}(t)}{R} \\
C\left(i \omega V(t)-i \omega V^{\prime}(t)\right) & =\frac{V^{\prime}(t)}{R} \\
R C \omega V(t) & =V^{\prime}(t)(R C \omega-i) \\
V^{\prime}(t) & =V(t) \frac{1}{1-\frac{i}{R C \omega}}
\end{aligned}
$$



Figure 2.9: High-pass RC filter

Alternatively, we can use the same voltage divider analysis from the series resistor case,

$$
\begin{aligned}
V^{\prime}(t) & =\frac{Z_{R}}{Z_{c}+Z_{R}} V(t) \\
& =\frac{R}{R-\frac{i}{\omega C}} V(t) \\
& =\frac{1}{1-\frac{i}{R C \omega}} V(t)
\end{aligned}
$$

which gives us the same expression.
We can now insert the expressions $V(t)=V_{\circ} e^{i \omega t}$ and $V^{\prime}(t)=V_{\circ}^{\prime} e^{i(\omega t+\delta)}$ where we have included a phase $\delta$ to account for any phase shift between the input and output.

$$
\begin{aligned}
V_{\circ}^{\prime} e^{i(\omega t+\delta)} & =V_{\circ} e^{i \omega t} \frac{1}{1-\frac{i}{R C \omega}} \\
V_{\circ}^{\prime} e^{i \delta} & =V_{\circ} \frac{1}{1-\frac{i}{R C \omega}} \\
\frac{V_{\circ}}{V_{\circ}^{\prime}} e^{-i \delta} & =1-\frac{i}{R C \omega}
\end{aligned}
$$

A minor amount of complex arithmetic is required to show that

$$
\begin{aligned}
V_{o}^{\prime} & =\frac{V_{o}}{\sqrt{1+\frac{1}{R^{2} C^{2} \omega^{2}}}} \\
\tan \delta & =\frac{1}{R C \omega}
\end{aligned}
$$

If we look at the limiting cases of high and low frequency, we see that at low frequency, $V^{\prime} \rightarrow 0$ while at high frequency $V^{\prime} \rightarrow V$. This is why the circuit in Figure 2.9 is called a highpass filter. It passes the high frequencies ( $\omega \gg 1 / R C$ ) while blocking the low frequencies
( $\omega \ll 1 / R C$ ). At high frequency, the phase shift goes to zero, while at low frequency the phase shift approaches $\pi / 2$ or 90 degrees.

The term $1 / R C$ is clearly the frequency which characterizes the filter, so we generally make the assignment

$$
\omega_{0}=\frac{1}{R C}
$$

which allows us to rewrite the ratio of output voltage to input voltage as

$$
\begin{equation*}
\frac{\left|V^{\prime}\right|}{|V|}=\frac{1}{\sqrt{1+\frac{\omega_{0}^{2}}{\omega^{2}}}} \tag{2.1}
\end{equation*}
$$

We note as well that at the characteristic frequency of the filter the phase shift is exactly $\pi / 4$.

### 2.5.2 The Low-Pass Filter



Figure 2.10: Low-pass RC filter

If we exchange the resistor and capacitor as in Figure 2.10, we obtain a low-pass filter. The circuit analysis can easily be performed using complex impedances

$$
\begin{aligned}
V^{\prime}(t) & =\frac{Z_{C}}{Z_{R}+Z_{C}} V(t) \\
& =\frac{\frac{-i}{\omega C}}{R-\frac{i}{\omega C}} V(t) \\
& =\frac{1}{1-i R C \omega} V(t)
\end{aligned}
$$

We can extract the limiting behavior by simply considering the impedance of the capacitor as a function of frequency. At low frequency, the capacitor's impedance is large so it
behaves a large resistance in the voltage divider. Thus we expect $V^{\prime} \rightarrow V$ as $\omega \rightarrow 0$, while at high frequencies the capacitor will turn into a short to ground so $V^{\prime} \rightarrow 0$.

Returning to the full expression for the low-pass voltage divider, some additional complex arithmetic yields

$$
\begin{aligned}
\frac{\left|V^{\prime}\right|}{|V|} & =\frac{1}{\sqrt{1+\omega^{2} R^{2} C^{2}}}=\frac{1}{\sqrt{1+\frac{\omega^{2}}{\omega_{0}^{2}}}} \\
\tan \delta & =-\omega R C=-\frac{\omega}{\omega_{0}}
\end{aligned}
$$

which has the inverse behavior as the high-pass case as expected. Note in particular that at low frequencies, where the filter has no effect on the amplitude, the phase shift is zero.

### 2.5.3 General Characteristics of Simple Filters and Amplifiers

Let us look in more detail at the implications of Eqn. (2.1) by plotting the attenuation as a function of frequency, using a $\log / \log$ plot where the $X$ axis is expressed in units of $\omega_{0}$ and the Y axis gives the decibels of attenuation as a function of the normalized frequency. Such a plot is called a Bode plot and the Bode plot for the high pass filter is shown in Figure 2.11.


Figure 2.11: Bode Plot of a high-pass RC filter.

The first characteristic of Figure 2.11 is that for $\omega \gg \omega_{0}$, there is no discernible attenuation visible - the filter does "nothing" to the signal, as we would hope. At $\omega=\omega_{0}$,

$$
\begin{aligned}
A & =20 \log _{10}\left(\frac{V^{\prime}}{V}\right) \\
& =20 \log _{10}\left(\frac{1}{\sqrt{1+1}}\right) \\
& =-3.01 \mathrm{~dB}
\end{aligned}
$$

so $\omega=\omega_{0}$ is often called "the 3 dB point" of a filter. As can be seen in the Bode plot, this is the frequency when the filter's response begins to change rapidly.

For lower frequencies, the response of the filter changes to a line with constant slope. We can easily derive this slope by using the Taylor expansion on the filter response.

$$
\begin{array}{rll}
\frac{V^{\prime}}{V} & =\frac{1}{\sqrt{1+\frac{\omega_{0}^{2}}{\omega^{2}}}} \\
& \approx & \frac{\omega}{\omega_{0}} \\
A\left(\omega \ll \omega_{0}\right) & = & 20 \log _{10}\left(\frac{\omega}{\omega_{0}}\right)
\end{array}
$$

Thus we see that when the frequency falls by a factor of ten, the attenuation increases by -20 dB . This fact is often expressed as "twenty dB per decade". For audio purposes, factors of two are quite important as a factor of two in frequency represents a musical interval of an octave. A factor of two decrease in $\omega$ will generate an increase in attenuation of $20 \log _{10} \frac{1}{2}=6 \mathrm{~dB}$, which is expressed as " 6 dB per octave".

The slope of $20 \mathrm{~dB} /$ decade is characteristic of first-order filters. It applies on the lowfrequency response of high-pass filters and the high-frequency response of low-pass filters. Stronger filters (with steeper slopes) can be built using combinations of resistors, capacitors, and inductors. The design of such filters is subject to a number of trade-offs and the researcher is recommended to consult a more detailed text if higher-order filters are needed.

### 2.5.4 RLC Circuits

High-pass and low-pass filters can be created using inductors instead of capacitors, but this is rarely done in practice since real inductors are challenging to work with. Inductors are used with capacitors, however, in more complex filters. Consider the circuit shown in Figure 2.12.

We can analyze this circuit by noticing that it is a parallel combination of an inductor and a capacitor in series with a resistor. We can begin by determining the complex impedance of the combined inductor-capacitor system.

$$
Z_{L C}=\frac{Z_{L} Z_{C}}{Z_{L}+Z_{c}}
$$



Figure 2.12: An RLC filter

$$
\begin{aligned}
& =\frac{(i L \omega)\left(\frac{-i}{C \omega}\right)}{i L \omega-\frac{i}{C \omega}} \\
& =-i \frac{\frac{L}{C}}{L \omega-\frac{1}{C \omega}} \\
& =\frac{-i L \omega}{L C \omega^{2}-1}
\end{aligned}
$$

Next, we can express the output voltage of the filter using the familiar series voltage divider formalism.

$$
\begin{aligned}
V^{\prime} & =V \frac{R}{R+Z_{L C}} \\
\frac{V^{\prime}}{V} & =\frac{1}{1-\frac{i L \omega}{R\left(L C \omega^{2}-1\right)}}
\end{aligned}
$$

At very low frequencies, the filter response is to pass the full signal. This can be understood intuitively as the capacitor blocking the signal, but the inductor allowing it to pass. At high frequencies, the filter similarly passes the full signal, which can be understood as the inductor blocking the signal but the capacitor passing it.

What about the behavior in between? In this case, neither the resistor nor the capacitor will be fully conducting. If we look carefully at the equation, we notice two characteristic frequencies in the denominator: $\omega_{L C}=\sqrt{\frac{1}{L C}}$, which is the characteristic frequency of the inductor/capacitor combination and $\omega_{L R}=\frac{R}{L}$, which is the characteristic frequency of a simple RL filter. We can re-write the attenuation equation using these characteristic
frequencies:

$$
\frac{\left|V^{\prime}\right|}{|V|}=\frac{1}{1-i \frac{\omega}{\omega_{L R}} \frac{\omega_{L C}^{2}}{\omega^{2}-\omega_{L C}^{2}}}
$$

In this form, we see that when $\omega \rightarrow \omega_{L C}$, the attenuation will become a maximum. In Bode plot format, as shown in Figure 2.13, it is clear why this filter is called a "notch filter": the filter takes a notch or chunk out of the frequency range. The width and maximum attenuation of the filter can be adjusted using $\omega_{L R}$, but not independently. Independent control of the width and maximum attenuation requires a higher-order filter (one with more components than a simple RLC circuit).


Figure 2.13: Bode plot for the notch filter

### 2.6 Real Components

### 2.6.1 Real Resistors

The resistors which you can pick up in the lab or purchase from supply stores have several imperfections. The first is that real resistors are manufactured to specific resistances within tolerances. Common tolerances are $\pm 5 \%$ and $\pm 1 \%$. Thus a $100 \Omega \pm 5 \%$ may actually be $96 \Omega$ or $103 \Omega$. Resistors with tolerances as tight as $0.01 \%$ can be obtained (for a price!), but generally circuits can be designed to work with $1 \%$ or $5 \%$ tolerances.

One could imagine going through a batch of $5 \%$ resistors and using an ohmmeter to find those within $0.1 \%$ of the desired resistance. However, even in this case the resistance may not be exactly as expected when the circuit is in use. This is because all resistors exhibit changes in resistance with changes in temperature. For relatively small changes in temperature relative to room temperature, the change is generally taken as linear so that a coefficient of temperature change can be defined as the fractional change in resistance per degree of temperature change.

$$
\frac{\Delta R}{R}=K \Delta T
$$

For most resistors, resistance increases with temperature. Typical temperature coefficients are in the range from $2.5 \times 10^{-5} \mathrm{~K}^{-1}$ to $5 \times 10^{-4} \mathrm{~K}^{-1}$.

### 2.6.2 Real Capacitors and Inductors

The capacitors which one can find in a lab fall into two large categories: small-signal capacitors and polarized power capacitors. Small-signal capacitors are appropriate for use in filters and other analog circuits and provide limited capacitance (up to $\approx 1 \mu \mathrm{~F}$ ). These capacitors can have a positive or negative voltage applied across their leads without any problem - as a result they are the appropriate choice for working with AC signals around ground. However, for the power supplies and other purposes, capacitors with higher capacitance are frequently needed. This capacitance is provided by devices with specialized electrolytic layers between the electrodes of capacitor. Generally these electrolytes work only for a defined potential direction between the electrodes, so these capacitors are described as polarized capacitors. Connecting a polarized capacitor in reverse is an effective way to generate a puff of smoke and a dead capacitor! Capacitors are also constructed with defined tolerances and exhibit temperature dependence. In general, the tolerances and temperature dependencies are much larger for the polarized capacitors compared to the small signal capacitors.

Inductors are generally built as coils of wire, sometimes wrapped around an iron core and sometimes with an air core. Since the wire involved has an inherent resistance, real inductors always have a resistive effect as well. As a result, inductors are usually modeled in realistic circuits as a inductor in series with a small value resistor. A realistic inductor complex impedance would then have the form

$$
Z_{L, \text { real }}=i \omega L+R_{L}
$$

so the true phase shift will depend on frequency. At high frequency, the inductor behavior will dominate the resistive behavior.

Inductors are generally considered to be bulky components compared to capacitors and resistors. In modern circuits, most inductors are quite small, such as the ferrite-bead inductors used to reduce high-frequency noise in the input stage of many circuits. Elsewhere, inductors have often been replaced by active circuits designed to perform the same function.

### 2.7 Diodes

Resistors, capacitors, and inductors are by themselves linear devices - the voltage across them is linearly proportional to current (resistors) or the integral of current (capacitors) or the derivative of current (inductors). By contrast, the simplest semiconductor device is the diode, which has a very nonlinear response to voltage. Diodes are crucial components for power supplies and conversion of AC signals into DC ones. They are also used in both the production and measurement of light, as is discussed further in Chapter 4.

The response of an ideal diode can be expressed as a piecewise resistance dependent on the voltage across the two leads of the device:

$$
R(V)=\left\{\begin{array}{cc}
0 & , V>0 \\
\infty & , V<0
\end{array}\right.
$$

An ideal diode thus provides no resistance to current flow in one direction, termed the forward direction, while blocking any flow in the opposite direction. In contrast to inductors, resistors, and (most) capacitors, the two leads of a diode are not equivalent - while a resistor or inductor can safely be plugged into a circuit two different ways, the polarity of the diode must be carefully observed for proper circuit behavior. The directionality of the diode's operation is reflected in the schematic symbol of the diode (Figure 2.14) where the triangle points in the forward direction of current flow (from the anode to the cathode).


Figure 2.14: Schematic symbol for a diode

Packaged diodes will generally have a band on one end to indicate the cathode or another indicator of the identity of the leads.

The characteristics described above are of course an idealization and we will discuss realistic diodes a bit later, but we can begin by looking at some simple circuits to understand the use of diodes.


Figure 2.15: A simple diode rectifier circuit (a) and $V(t)$ behavior (b).

### 2.7.1 Diode Circuits

Consider the circuit in Figure 2.15a. This circuit is another form of the voltage divider circuit, with the diode now playing the role of $R_{1}$. If we feed in a time-varying voltage $V(t)=V_{\circ} \sin \omega t$, what will the voltage at the output $\left(V^{\prime}(t)^{\prime}\right)$ be?

When the $V(t)>0$, the resistance of our ideal diode will be zero, so

$$
V^{\prime}(t)=\frac{R}{R+0} V(t)=V(t) .
$$

On the other hand, when $V(t)<0$ the diode's resistance will become infinite so

$$
V^{\prime}(t)=\frac{R}{R+\infty} V(t)=0 .
$$

The resulting waveform is shown by the solid curve in Figure 2.15b. The dotted portion of the waveform has been "chopped off" by the diode circuit. This circuit shows the use of a diode as a rectifier. Rectifiers are important for the conversion of alternating current into direct current. In Figure 2.15a, the resistor stands for a generic load which will see an oscillating but always positive input voltage.

The variation of the voltage across the load or droop can be reduced by adding a capacitor between the output of the diode and ground as shown in Figure 2.16a. The capacitor serves as a charge source during the period when the diode is reverse-biased and not conducting.

The capacitor will follow the RC transient behavior described above with

$$
V_{c}(t)=V_{\circ}\left(1-e^{-t / R C}\right)
$$

. Thus if $\omega<\frac{1}{R C}$ the capacitor will not significantly discharge before the diode returns to forward conduction and is able to recharge the capacitor as seen in Figure 2.16b. The resulting waveform is much closer to a true constant voltage than the circuit without the capacitor. The circuit is a type of low-pass filter on the incoming signal and the large-value (usually polarized) capacitors which are found in power supplies are generally referred to as filter capacitors.


Figure 2.16: A diode rectifier circuit including a filter capacitor (a) and the resulting $V(t)$ behavior (b).

### 2.7.2 Realistic Diode Characteristics

Real diodes are not able to achieve quite the conversion between zero and infinite resistance which is implied by the ideal diode model. A real diode has a dependency of current on voltage similar to that in Figure 2.17. This sort of chart is known as an "IV curve" and the curve for a real diode shows a number of interesting features.


Figure 2.17: Characteristic diode current versus voltage behavior (IV curve). The magnitude of $I_{r l}$ is generally $\mathrm{O}(10 \mathrm{nA})$ or smaller.

The first feature of a realistic diode is that the resistance of the device does not go to zero immediately for positive voltages. Instead, there is a region where the diode is roughly linear $(I \propto V)$ before the curve turns sharply upwards in a roughly quadratic rise
$\left(I \propto V^{2}\right)$. The amount of voltage required to get to the vertical portion of the diode curve is called the forward diode drop, and depends on the technology of the diode in question. For a silicon diode, the forward drop is $\approx 0.7 \mathrm{~V}$. From a fundamental device-physics point of view, the forward diode drop is related to the gap between the valence and conduction bands of the semiconductor material used to construct the diode. We will discuss these details further in Chapter 4 when we study the device physics relevant to photodiodes and other semiconductor sensors.

On the negative bias side, the diode exhibits a nearly constant reverse leakage current $I_{r l}$ over a wide range of voltages. This leakage current is generally 10 nA or smaller. At a large reverse voltage $\left(V_{r b}\right)$ the diode goes into reverse breakdown and the reverse current rises very rapidly with voltage. At this point, the diode is not useful as a rectifier and a current-limiting resistor must be included in series to avoid damaging the diode. If the current is safely limited, however, reverse breakdown is not a damaging condition for most diodes. In fact, one can use the fact that the reverse breakdown slope is much steeper than the forward slope to perform voltage regulation.

Zener diodes are a class of diodes with carefully engineered reverse breakdown voltages which can be used for simple voltage regulation. Figure 2.18 shows a Zener regulator circuit consisting of a bias resistor $R_{B}$ in series with the Zener diode which supplies the load $R_{L}$. Notice that the Zener diode is installed "in reverse" to make use of the sharp reverse breakdown transition. The transition is sufficiently sharp that the Zener will sink enough current to make the voltage across it equal to $V_{r b}$. This regulated constant voltage is then available to the load. A Zener regulator can be combined with the filtered rectifier described above to provide a constant DC voltage source relatively free from ripple.


Figure 2.18: A simple Zener-diode-based voltage regulator.

## Problems

1. Determine the Thevenin equivalent voltage and resistance for the circuit shown below.

2. Determine if the circuit shown below is a low-pass or high-pass filter, and find the capacitance needed to obtain the same $\omega_{0}$ for a capacitor-based filter.

3. Suppose one has a signal with a frequency of 1 kHz which is being swamped by a 60 Hz background noise. One solution might be to apply a simple high-pass RC filter. If the filter is designed with $\omega_{0}=1 \mathrm{kHz}$, by what factor will the 60 Hz noise be reduced by the filter?
4. Determine the output of the filter below as a function of frequency.

5. The circuit shown below uses a special type of diode called an LED which produces light when a current flows through it (these devices are discussed in more detail in Chapter 4). Such a diode exhibits a large forward diode drop which depends on the color of the LED. Given the LED's forward bias voltage ( $V_{f b}=2.3 \mathrm{~V}$ ), determine the correct resistor to use if the LED should have 10 mA of current flowing through it.

6. Sketch $V^{\prime}(t)$ for the circuit below with the given input waveform $V(t)$. Assume an ideal diode (no forward diode drop and infinite reverse breakdown voltage).

7. Determine the minimum filter capacitance required to keep the voltage droop below 5 mV for a simple single diode rectifying a $10 \mathrm{~V}, 60 \mathrm{~Hz}$ input and feeding a $Z_{L}=35 \Omega$ load as in Fig. 2.16(a).

## Chapter 3

## Operational Amplifiers

Physical measurements generally begin with a sensor which provides a voltage or current signal. Frequently, the signals from sensors are quite small and many sensors exhibit high output impedance which limits the ability to use signals for plotting and display. Also, the interesting physical measurement may not be a single voltage but rather a difference between two signals or a sum. We can use amplifiers to increase the amplitude of signals, but also to perform other more sophisticated signal processing.

### 3.1 Amplification Principles

The basic role of an amplifier is to make a signal larger, either in voltage or in current. In either case, we can define the gain as the ratio between the output and input amplitudes.

$$
G=\frac{A_{o}}{A_{i}}
$$

The amplitudes will generally be the input and output voltages, but could also be the input and output powers or currents. In fact, as we will see below, one of main uses of op-amps is to buffer signals - to provide a power boost with no change in amplitude.

The first law of thermodynamics demands that if the amplifier is to provide more power than it absorbs from its input, it must make up the difference somehow. Thus, the amplifier must be connected to a power supply to provide the necessary energy. In fact, most amplifiers need to be connected to two power supplies. Sometimes a single power supply will be used and the second supply will be taken as ground, but usually a separate positive and negative supply will be used as seen in Figure 3.1.

One might wonder why separate supplies are so often used - adding a second power supply increases the complexity and cost of a system. The main reason is that amplifiers cannot amplify signals beyond the range of their power supplies. For a small AC signal hovering near ground, a single-supply setup cannot be used - the negative side excursions will necessarily be clipped. In fact, most amplifiers require some "head room" between the desired output voltage and the power supplies. The maximum and minimum allowable output voltages are called the saturation voltages ( $S_{+}$and $S_{-}$). Safe design of an amplifier


Figure 3.1: Schematic symbol for an operational amplifier showing the power supply terminals.
circuit requires keeping the output voltage within this range, and thus limits the range of the input voltage.

$$
\frac{S_{-}}{G} \leq V_{i} \leq \frac{S_{+}}{G}
$$

### 3.2 Ideal Operational Amplifiers

The most commonly-used building block amplifier is the operational amplifier or op-amp. The schematic symbol for an op-amp is shown in Figure 3.1. This version of the symbol includes the power supply connections which are often omitted for simplicity - we will follow this convention for the most of the rest of the circuits in this chapter, except when the power supply connections are relevant to the discussion.

The op-amp is a three-terminal device with one output terminal and two input terminals. The two input terminals are called the "inverting input" and the "non-inverting input" and are labeled with a minus-sign and plus-sign respectively. The output of the op-amp is proportional to the voltage difference between the non-inverting and inverting inputs.

$$
V_{o}=G_{o}\left(V_{+}-V_{-}\right)
$$

Here we have written the op-amp gain as $G_{o}$ because it is the so-called "open-loop" gain.
For an ideal op-amp, three important assumptions can be made.

1. The inputs of an ideal op-amp draw no current, so $I_{+}=0$ and $I_{-}=0$. This is equivalent to stating that the impedance of an op-amp input is infinite ${ }^{1}$.
2. The output impedance of an ideal op-amp is zero - the op-amp can drive any required current.
3. The open-loop gain of an ideal op-amp is very large $\left(G_{o} \rightarrow \infty\right)$.

[^0]

Figure 3.2: The inverting amplifier.

Since the open-loop gain of an ideal op-amp is infinite, it does not make sense to use it to simply directly amplify the difference the two inputs - the amplifier will immediately enter saturation on either the positive or negative side if $V_{+} \neq V_{-}$. Instead, the op-amp is generally used in a closed-loop configuration where the output is looped back in some way to the inverting input to provide negative feedback stabilization.

### 3.2.1 The Inverting Amplifier

## Ideal Analysis

To understand how closing the loop provides significant benefits, let us analyze the circuit in Figure 3.2. This circuit is called an inverting amplifier, for reasons which will become clear.

The non-inverting input of the op-amp is connected directly to ground, so the relationship between $V_{\text {out }}$ and $V_{-}$is quite simple.

$$
V_{\text {out }}=G_{o}\left(V_{+}-V_{-}\right)=-G_{o} V_{-}
$$

For the analysis of the circuit, it will be more convenient to work with $V_{\text {out }}$ than $V_{-}$, so we will use this relationship to make the necessary replacement.

Since we know the voltages at both ends of $R_{2}$ we can easily write down the current across this resistor.

$$
\begin{aligned}
I_{2} & =\frac{V_{-}-V_{\text {out }}}{R_{2}} \\
& =\frac{\frac{-V_{\text {out }}}{G_{o}}-V_{\text {out }}}{R_{2}} \\
& =\frac{-V_{\text {out }}}{R_{2}}\left(1+\frac{1}{G_{o}}\right)
\end{aligned}
$$

Similarly, we can write down the current through $R_{1}$.

$$
I_{1}=\frac{V_{\text {in }}-V_{-}}{R_{1}}=\frac{V_{\text {in }}+\frac{V_{\text {out }}}{G_{o}}}{R_{1}}
$$

Now we will invoke the first principle of the ideal op-amp, which states that no current will flow into or out of an op-amp's input. Applying Kirchoff's current sum rule, we see that

$$
I_{1}=I_{2}
$$

We can then substitute the values obtained above for the two currents.

$$
\begin{equation*}
\frac{V_{\text {in }}+\frac{V_{\text {out }}}{G_{o}}}{R_{1}}=\frac{-V_{\text {out }}}{R_{2}}\left(1+\frac{1}{G_{o}}\right) \tag{3.1}
\end{equation*}
$$

We can take $G_{o} \rightarrow \infty$ in (Eq. 3.1) to find the closed-loop gain of the circuit in the ideal op-amp case.

$$
\begin{aligned}
\frac{V_{\text {in }}}{R_{1}} & =\frac{-V_{\text {out }}}{R_{2}} \\
V_{\text {out }} & =-V_{\text {in }} \frac{R_{2}}{R_{1}} \\
G \equiv \frac{V_{\text {out }}}{V_{\text {in }}} & =\frac{-R_{2}}{R_{1}}
\end{aligned}
$$

Thus we see that the output of the circuit is indeed a voltage-amplified version of the input, with the polarity of the signal switched. The amount of amplification is set not by any inherent characteristic of the op-amp but rather by the external resistors. A single variety of op-amp can thus be used for a wide array of circuits - a characteristic which leads to mass production and low cost for an individual op-amp. We also note that when $G_{o} \rightarrow \infty$, $V_{-} \rightarrow 0=V_{+}$. This is a common characteristic of closed-loop op-amp circuits which we discuss in more detail below.

If we intend to use this circuit to amplify signals from a high-impedance sensor, we should evaluate the input impedance and output impedance of the circuit. The output impedance of the circuit is straightforward; according to the third principle of ideal opamps, the output impedance of the circuit will be zero as the output is directly driven by the zero-impedance output of the op-amp. The input impedance can be evaluated by considering the current drawn by the circuit for a given input voltage $V_{i n}$. This will simply be the current through $R_{1}$ which we have already calculated. In the case where $G_{o} \rightarrow \infty$,

$$
I_{1}=\frac{V_{i n}}{R_{1}}
$$

so the input impedance of the inverting amplifier is simply set by the value of $R_{1}$.

## Effect of Finite Open-Loop Gain

In fact, the open-loop gain of real op-amps is large $\left(\mathrm{O}\left(10^{5}\right)\right.$ ), but not infinite. What effect does finite open-loop gain have on the inverting amplifier? We can reorganize (Eq. 3.1) to obtain

$$
\begin{align*}
\frac{V_{\text {in }}}{R_{1}} & =\frac{-V_{\text {out }}}{R_{2}}\left(1+\frac{1}{G_{o}}+\frac{R_{2}}{R_{1} G_{o}}\right) \\
\frac{V_{\text {out }}}{V_{\text {in }}} & =\frac{-R_{2}}{R_{1}} \cdot \frac{1}{1+\frac{1}{G_{o}}+\frac{R_{2}}{R_{1} G_{o}}} \\
G & =\frac{-R_{2}}{R_{1}} \cdot \frac{1}{1+\left(1+\frac{R_{2}}{R_{1}}\right) / G_{o}} \tag{3.2}
\end{align*}
$$

The effect of finite open-loop gain is thus to reduce the true output gain by an amount which depends on the designed closed-loop gain. To keep the op-amp behavior ideal, we must require that

$$
1+\frac{R_{2}}{R_{1}}=1-G \ll G_{o}
$$

This requirement generally limits the acceptable closed-loop gain to at most a factor of 1000, but one chain multiple stages of amplifiers to achieve higher total gain if necessary.

### 3.2.2 The Non-inverting Amplifier



Figure 3.3: The non-inverting amplifier.

With the inverting amplifier behind us, let us shift our analysis to an amplifier design where the input is connected to the non-inverting input of the op-amp rather than the inverting input - the non-inverting amplifier (Figure 3.3).

In the non-inverting amplifier, $V_{i n}=V_{+}$so we can write the output voltage as

$$
V_{\text {out }}=G_{o}\left(V_{\text {in }}-V_{-}\right)
$$

The currents across the two resistors are also simple to write down.

$$
\begin{aligned}
& I_{1}=\frac{V_{-}}{R_{1}} \\
& I_{2}=\frac{V_{\text {out }}-V_{-}}{R_{2}}
\end{aligned}
$$

We can again apply the first principle of ideal op-amp behavior and Kirchhoff's current sum rule to obtain the relationship between the input and output voltages.

$$
\begin{align*}
\frac{V_{-}}{R_{1}} & =\frac{V_{\text {out }}-V_{-}}{R_{2}} \\
V_{-}\left(\frac{1}{R_{1}}+\frac{1}{R_{2}}\right) & =\frac{V_{\text {out }}}{R_{2}}  \tag{3.3}\\
\left(V_{\text {in }}-\frac{V_{\text {out }}}{G_{o}}\right) \cdot\left(\frac{1}{R_{1}}+\frac{1}{R_{2}}\right) & =\frac{V_{\text {out }}}{R_{2}}
\end{align*}
$$

When we again let $G_{o} \rightarrow \infty$, we can determine the closed-loop gain.

$$
G \equiv \frac{V_{\text {out }}}{V_{\text {in }}}=1+\frac{R_{2}}{R_{1}}
$$

In this case the gain is positive, so an incoming signal will not be inverted. The noninverting configuration is capable of gains $\geq 1$, while the inverting amplifier is also able to reduce the amplitude of signal - an achievement one could make passively using a voltage divider!

The non-inverting configuration is particularly useful for small voltage signals given that its input impedance is infinite in the ideal case (and often in the $G \Omega$ range for real op-amps). A particularly simple use of the non-inverting configuration is the buffer shown in Figure 3.4. The buffer is a non-inverting amplifier with $R_{1}=\infty$ and $R_{2}=0$ which gives a gain of 1 . The buffer converts a possibly high-impedance signal into a low-impedance one, providing isolation between a sensor or signal and further signal processing steps or output.


Figure 3.4: The op-amp buffer or unity-gain amplifier.

### 3.2.3 The Virtual Short Circuit

Let us determine what the actual value of $V_{-}$is from the analysis of the non-inverting amplifier, using as the simplest starting point (Eq. 3.3).

$$
\begin{aligned}
V_{-}\left(1+\frac{R_{2}}{R_{1}}\right) & =V_{\text {out }} \\
& =G_{o}\left(V_{+}-V_{-}\right) \\
V_{-}\left(1+\frac{1}{G_{o}}+\frac{R_{2}}{R_{1} G_{o}}\right) & =V_{+}
\end{aligned}
$$

Thus, as $G_{o} \rightarrow \infty$, we see that $V_{-} \rightarrow V_{+}$just as it did for inverting amplifier case.
The equality of voltage on the two inputs of an op-amp is a general characteristic of stable op-amp circuits. This behavior is commonly described as a virtual short circuit between the two inputs. In the case where one of the inputs is tied to ground, the other is often referred to as a virtual ground. Through feedback, the large gain of the op-amp will cause the voltage on the inverting input to match the non-inverting input even if the noninverting input's voltage is changing. This behavior is very useful for circuit analysis, as we can use the virtual short to simplify the calculation. One should be clear, however, that the short circuit is virtual - a circuit design should not actually tie the inputs of the op-amp together!

## Example



Figure 3.5: A meter amplifier.

Let us apply the principle of the virtual short circuit to analyze the meter control circuit above. Needle meters are used for displaying all sorts of values, and generally have an angular displacement which is proportional to the current flowing through them. However, the resistance of the needle meter is not necessarily constant as a function of angular displacement. This makes it difficult to use the meter to read voltage directly or by including a resistor in series with the meter. The circuit in Figure 3.5 is designed to correct for the meter's varying resistance.

Let us analyze this circuit using the principle of the virtual short circuit. In this case, we see that

$$
V_{+}=V_{i n}=V_{-}
$$

and thus the current through the resistor is simply

$$
I=\frac{V_{-}}{R}=\frac{V_{i n}}{R} .
$$

However, because the inputs of an op-amp do not draw current, this must also be the current through the meter regardless of the resistance of the meter. The circuit automatically compensates for the changing resistance of the meter as a function of current. Of course, this compensation will only work while the required drive voltage does not drive the opamp into saturation. The circuit also makes it simple to match the dynamic range of the meter with the desired voltage range - the resistor can be tuned as necessary to match the full-deflection current of the meter.

### 3.3 Non-Ideal Op Amps

For the ideal op-amp, the inputs do not draw or source currents and the open-loop $V_{\text {out }}$ will be zero when $V_{-}=V_{+}$. Unfortunately, these two assumptions do not exactly hold for real op-amps, which can have a number of negative effects. The small currents which do flow into or out of the op-amp's inputs are referred to as bias currents, while the differential input voltage which does result in a zero open-loop output is referred to as the offset voltage. Opamps also have some tendency to amplify the sum of $V_{-}$and $V_{+}$which is called commonmode amplification. Finally, op-amps have maximum frequencies which they are able to amplify successfully and these limits depend on the closed-loop gain as well as the signal amplitude. All of these deviations from the ideal op-amp model will have implications for the performance of real amplifiers and are particularly important for very high frequency and high gain designs.

### 3.3.1 Offset Voltage

The offset voltage can be modeled by placing a small voltage supply $\left(V_{o s}\right)$ in series with the positive input of an ideal op-amp as in Figure 3.6. The sign of the offset voltage varies from device to device: it depends on small differences in the input stages of the inverting and non-inverting inputs. The magnitude of the offset voltage depends on device technology ("bipolar" op-amps tend to have smaller $V_{o s}$ than "FET" op-amps), but is generally in the range $100 \mu \mathrm{~V}-10 \mathrm{mV}$.

We can easily see the effect of $V_{o s}$ by studying a simple inverting amplifier with a large gain (say 1000x). For the realistic op-map we will observe that $V_{-}$will not go to ground as we would expect from the principle of the virtual short-circuit. Instead, it will approach $V_{\text {os }}$. We can evaluate the behavior of the circuit using Kirchhoff's current rule across the feedback voltage divider of the inverting amplifier.


Figure 3.6: A model of offset voltage.

$$
\begin{aligned}
I_{1} & =I_{2} \\
\frac{V_{\text {in }}-V_{o s}}{R_{1}} & =\frac{V_{\text {os }}-V_{\text {out }}}{R_{2}} \\
V_{\text {out }} & =\frac{-R_{2}}{R_{1}}\left(V_{\text {in }}-V_{o s}\right)+V_{o s}
\end{aligned}
$$

For a 1000 x amplifier, the output voltage for a zero input voltage will thus be $1000 V_{o s}$ : The amplifier will directly amplify $V_{o s}$. For an op-amp with $V_{o s}=5 \mathrm{mV}$, $V_{o u t}$ will be 5 V with no input voltage applied!

There are a number of ways to limit the effect of $V_{o s}$ in the relatively few cases where it matters. One is to convert the inverting amplifier into a summing amplifier as described below and provide a variable resistor to trim the output of the amplifier to zero when the input is grounded. Some op-amps even provide special inputs to allow trimming of the offset voltage without connecting to the inputs. Unfortunately, $V_{o s}$ is generally a function of temperature, so one cannot set the resistor once and forget it - it may require adjustment with temperature. For applications where DC amplification is not needed, a different solution can be used which is described in Problem 1 in this chapter.

### 3.3.2 Bias Current

While ideal op-amps do not draw any current through their inputs, real op-amps require a small current to bias their input stages. The magnitude of this current depends on the technology in use. In this case, the "FET" op-amps have much smaller bias currents ( $\sim$ $1 \mathrm{pA})$ than the typical "bipolar" op-amp $(\sim 1 \mu \mathrm{~A})$. The bias current for each input will be different, but generally quite similar : $\frac{I_{b+}-I_{b-}}{I_{b+}+I_{b-}} \ll 1$. For circuits which involve large-value resistors, these bias currents can have significant effects as the currents involved become similar in magnitude to the signal-induced currents.

To study this, consider an inverting amplifier with $R_{1}=500 \mathrm{k} \Omega$ and $R_{2}=5 \mathrm{M} \Omega$. The large value of $R_{1}$ is chosen to raise the input impedance of the amplifier. However, it necessarily implies a small current flowing through the system. To apply Kirchhoff's current rule, we must include the bias current $I_{b-}$ as well. We will still assume that $V_{-}=V_{+}=0$, however.

$$
\begin{aligned}
I_{1} & =I_{2}+I_{b-} \\
\frac{V_{\text {in }}}{R_{1}} & =\frac{-V_{\text {out }}}{R_{2}}+I_{b-} \\
V_{\text {out }} & =\frac{-R_{2}}{R_{1}} V_{\text {in }}+R_{2} I_{b-}
\end{aligned}
$$

Thus, if $I_{b-}=100 \mathrm{nA}$, the actual output voltage will be smaller than the ideal by 500 mV ! Thus, the bias current effectively limits the size of $R_{2}$ which can be used which will limit the gain which can be achieved with a fixed input impedance. Notice that we take a current flowing into the input as positive.


Figure 3.7: Cancellation of bias current effects in the inverting amplifier.

Interestingly, most of the effect of bias current can be canceled in an inverting amplifier by adding a resistor between ground and the non-inverting input as seen in Figure 3.7. For the ideal op-amp, the effect of $R_{3}$ is completely negligible: since no current flows through the input, there will be no voltage drop across the resistor. However, the value of $R_{3}$ appears for the realistic op-amp. In particular, op-amps are often constructed with matched input stages so that the bias currents for the inputs are very similar to each other so that we can assume $I_{b-}=I_{b+}=I_{b}$. The voltage at the non-inverting input will no longer be zero, but rather $-I_{b} R_{3}$. This changes the virtual short-circuit voltage on $V_{-}$as well, so we must re-evaluate the current flow calculation above.

$$
I_{1}=I_{2}+I_{b}
$$

$$
\begin{aligned}
\frac{V_{\text {in }}+I_{b} R_{3}}{R_{1}} & =\frac{-I_{b} R_{3}-V_{\text {out }}}{R_{2}}+I_{b} \\
V_{\text {out }} & =\frac{-R_{2}}{R_{1}} V_{\text {in }}+I_{b}\left(R_{2}-\frac{R_{2} R_{3}}{R_{1}}-R_{3}\right)
\end{aligned}
$$

We can choose $R_{3}$ to cancel the effect of $I_{b}$ by setting the term in parenthesis to zero. When we do so, we find that

$$
R_{3}=\frac{R_{1} R_{2}}{R_{1}+R_{2}}
$$

which is simply the parallel equivalent resistance of $R_{1}$ and $R_{2}$. This technique will mostly cancel the effects of bias current for the inverting amplifier, leaving only the effects which come from differences in the bias current for the two inputs.

### 3.3.3 Common-Mode Amplification

In open-loop configuration, an ideal op-amp amplifies only the voltage difference between the inputs - a common shift in voltage should not be amplified. These conditions can be expressed algebraically as

$$
V_{\text {out }}=G_{o}\left(V_{+}-V_{-}\right)+G_{c m}\left(\frac{V_{+}+V_{-}}{2}\right)
$$

where for a ideal op-amp $G_{c m}=0$. The gain $G_{c m}$ is referred to as the common mode gain. Generally, the common mode gain is compared to the differential gain as the common mode rejection ratio (CMRR) which is often expressed in decibels:

$$
\mathrm{CMRR}=20 \log _{10}\left(\frac{\left|G_{o}\right|}{\left|G_{c m}\right|}\right)
$$

Typical CMRR values for op-amps are in the range 70 dB to 120 dB or more.
Common-mode amplification is not a problem for the inverting amplifier - since both inputs are close to ground, the common mode voltage is small. For the non-inverting amplifier, however, the effect of common-mode amplification can be significant since both inputs are similar in value and often far from ground. Common-mode rejection is a generalpurpose concept which we will observe again when discussing the instrumentation amplifier.

### 3.3.4 Frequency Response

The open-loop gain of op-amps is not constant with frequency. In fact, most op-amps exhibit a decrease in gain identical to a low-pass RC filter - a $20 \mathrm{~dB} /$ decade decline from quite a low frequency (often in $100 \mathrm{~Hz}-1 \mathrm{kHz}$ range). This RC-like decrease in gain is not by chance - most op-amps contain an RC-network whose roll is to provide compensation of the op-amp to avoid it falling into self-driven oscillation. Op-amps are characterized by the frequency at which the open-loop gain falls to unity which is often called either the unitygain bandwidth or the gain-bandwidth product. This parameter $\left(f_{t}=\omega_{t} / 2 \pi\right)$ is specified on op-amp datasheets and is important even for closed loop designs.

To understand why the open-loop unity-gain bandwidth is important to closed-loop amplifiers, recall (Eq 3.2) which gives the inverting amplifier gain including finite open-loop gain effects. For the compensated op-amp with a unity-gain bandwidth $\omega_{t}$, we can write down the frequency-dependent gain using the low-pass filter behavior from Chapter 2.

$$
G_{o}(\omega)=\frac{1}{\sqrt{1+\frac{\omega^{2}}{\omega_{t}^{2}}}} \approx \frac{\omega_{t}}{\omega}
$$

Thus, we can re-write (Eq 3.2) as

$$
G(\omega)=\frac{-R_{2}}{R_{1}} \cdot \frac{1}{1+\left(1+\frac{R_{2}}{R_{1}}\right) \frac{\omega}{\omega_{t}}}
$$

The gain of the inverting amplifier thus appears similar to a low-pass filter. If we expand this for large frequencies, we find

$$
G(\omega) \approx \frac{\frac{\omega_{t}}{1+\frac{R_{2}}{R_{1}}}}{\omega}
$$

We can then define the characteristic closed-loop 3dB point

$$
\omega_{3 \mathrm{~dB}}=\frac{\omega_{t}}{1+\frac{R_{2}}{R_{1}}}=\frac{\omega_{t}}{1-G_{d c}}
$$

where $G_{d c}$ is the DC gain of the amplifier. The gain of the inverting amplifier is constant for $\omega \ll \omega_{3 \mathrm{~dB}}$ and falls at $-20 \mathrm{~dB} /$ decade above this critical point. As the closed-loop gain of the amplifier increases, the critical frequency $\omega_{3 \mathrm{~dB}}$ falls. This trade-off between frequency (or bandwidth) and gain is the reason for calling $\omega_{t}$ the gain-bandwidth product.

## Example

Let us find $f_{3 \mathrm{~dB}}$ and the gain at 10 kHz for the inverting amplifiers constructed using an op-amp with $f_{t}=2.0 \mathrm{MHz}$ and with DC gains of $-1 \mathrm{X},-10 \mathrm{X},-100 \mathrm{X}$ and -1000 X .

We can find the 3 dB frequency by simply using the relationship between $f$ and $\omega$ and the definition of $\omega_{3 \mathrm{~dB}}$.

$$
f_{3 \mathrm{~dB}}=\frac{1}{2 \pi} \frac{2 \pi f_{t}}{1-G_{c l}}=\frac{f_{t}}{1-G_{d c}}
$$

We can then re-write (Eq 3.2) simply as

$$
\begin{aligned}
G(\omega) & =G_{d c} \cdot \frac{1}{1+\left(1-G_{d c}\right) \frac{f}{f_{t}}} \\
& =G_{d c} \cdot \frac{1}{1+\frac{f}{f_{3 \mathrm{~dB}}}}
\end{aligned}
$$

| DC Gain | $f_{3 \mathrm{~dB}}$ | Gain at 10 kHz |
| :--- | :--- | :--- |
| -1 X | 2.0 MHz | -0.995 X |
| -10 X | 180 kHz | -9.47 X |
| -100 X | 20 kHz | -66.7 X |
| -1000 X | 2 kHz | -167 X |

### 3.3.5 Slew Rate

The frequency-gain limitations which arise from the limited bandwidth of op-amps are not the only limitations on the amplitude/frequency response of op-amp amplifiers. Opamps are also limited in the rate at which their outputs can change. This limitation applies independent of the unity-gain bandwidth limit and is called the slew rate limitation. Slew rates are specified in units of $\mathrm{V} / \mu \mathrm{s}$ with values as small as $1 \mathrm{mV} / \mu \mathrm{s}$ and as large as $5 \mathrm{kV} / \mu \mathrm{s}$. Slew rate limitations cause an op-amp circuit to fail to respond to large signals as rapidly as small signals at the same frequencies. The result of slew rate limitation is frequently a distortion of the wave shape - rapid voltage shifts become rounded as the output is limited a linear ramp at the maximum slew rate.

(a)

(b)

Figure 3.8: Waveform distortions induced by slew limits. In both figures, the dotted lines show the expected response from the circuit in the absence of slew limits while solid curves show the actual response.

Two examples of waveforms which have been distorted by slew rate limits are shown in Figure 3.8. In Fig 3.8a, the expected output is a step function. The output is not able to reproduce this because of both the finite gain-bandwidth product of the amplifier and the slew rate limitation. In this case, the slew rate limitation is dominant as evidenced by the linear voltage increase at the maximum slew rate of the op-amp. If the circuit was limited by the gain-bandwidth product, one would expect an exponential behavior depending on $\omega_{t}$ as for the RC transients discussed in Section 2.4. In Fig 3.8b, the expected output is a sinusoidal wave, but the slew-rate limitation produces a triangle wave as the amplifier lags behind the desired signal. In both cases, the linear ramp distortion of the waveform is the hallmark of slew-rate limitation.

### 3.4 Op-Amp Applications

Now that we have studied the two main op-amp circuits in both the ideal and non-ideal case, we can proceed to look at a number of interesting applications for this flexible and powerful element.

### 3.4.1 Summing and Differencing Amplifiers

## Summing Amplifier



Figure 3.9: The summing or averaging amplifier.

Figure 3.9 shows the summing amplifier, which is derived from the inverting amplifier by adding additional resistors in parallel to the original $R_{1}$. We can most easily analyze this circuit by using the virtual ground at $V_{-}$. Thus, the current across each of the input resistors will be simply related to the voltage on that input : $I_{1}=\frac{V_{1}}{R_{1}}$ and so on. However, since the input does not draw current, we see that the current across $R_{f}$ must be the sum of the currents across the input resistors.

$$
\begin{aligned}
I_{f} & =I_{1}+I_{2}+I_{3} \\
\frac{-V_{\text {out }}}{R_{f}} & =\frac{V_{1}}{R_{1}}+\frac{V_{2}}{R_{2}}+\frac{V_{3}}{R_{3}} \\
V_{\text {out }} & =-R_{f}\left(\frac{V_{1}}{R_{1}}+\frac{V_{2}}{R_{2}}+\frac{V_{3}}{R_{3}}\right)
\end{aligned}
$$

Thus, the output voltage is the weighted sum of the input voltages. We can vary the weights to perform different calculations. In particular, if we take $R_{1}=R_{2}=R$ and $R_{f}=R / 3$ then
the output voltage will be the negative average of the input voltages.

$$
\begin{aligned}
V_{\text {out }} & =-\frac{R}{3}\left(\frac{V_{1}}{R}+\frac{V_{2}}{R}+\frac{V_{3}}{R}\right) \\
& =-\frac{V_{1}+V_{2}+V_{3}}{3}
\end{aligned}
$$

## Differencing Amplifier

Since we cannot pull a negative resistor off the shelf, however, we can only perform sums with positive weights using the summing amplifier. To perform a differencing operation, we can use the circuit in Figure 3.10, which is a differencing amplifier.


Figure 3.10: The differencing amplifier.

We can immediately derive the expressions for the voltages at the two op-amp inputs in terms of the two input voltages and the op-amp output voltage.

$$
\begin{aligned}
V_{+} & =V_{b} \frac{R_{4}}{R_{3}+R_{4}} \\
I_{1} & =I_{2} \\
\frac{V_{a}-V_{-}}{R_{1}} & =\frac{V_{-}-V_{\text {out }}}{R_{2}}
\end{aligned}
$$

This analysis of this circuit is greatly simplified by the assumption that the negative feedback is successful and that the principle of the virtual short circuit applies so that
$V_{-}=V_{+}$. In this case, we can solve for the output voltage in terms of the input voltages.

$$
\begin{aligned}
V_{\text {out }} & =V_{-}\left(1+\frac{R_{2}}{R_{1}}\right)-\frac{R_{2}}{R_{1}} V_{a} \\
& =V_{b} \frac{R_{4}}{R_{3}+R_{4}} \frac{R_{1}+R_{2}}{R_{1}}-\frac{R_{2}}{R_{1}} V_{a} \\
& =V_{b} \frac{R_{1}+R_{2}}{R_{3}+R_{4}} \frac{R_{4}}{R_{1}}-\frac{R_{2}}{R_{1}} V_{a}
\end{aligned}
$$

This arrangement allows for relative weighting of the two inputs. If we wish to make a direct subtraction (as is often the case), we can set $R_{3}=R_{1}$ and $R_{4}=R_{2}$. Then we obtain

$$
V_{\text {out }}=\frac{R_{2}}{R_{1}}\left(V_{b}-V_{a}\right)
$$

This circuit performs the direct subtraction exactly as we would want and the differential gain can be controlled by the resistor ratio.

### 3.4.2 The Instrumentation Amplifier

The simple differencing amplifier, while useful for performing differencing operations, has difficulties in the case of small signals, particularly for small signals on top of a large constant voltage. One obvious area of concern is the input impedance of the differencing amplifier - for small precise signals, we want a large constant impedance. The input impedance for $V_{b}$ is clearly $R_{3}+R_{4}$. The input impedance for $V_{a}$ is actually non-constant and depends on $V_{b}$. An obvious improvement to the simple differential amplifier is thus to add non-inverting amplifier stages to both buffer both inputs - raising the input impedance to a large value at the cost of adding two op-amps to the design. These amplifiers can be constructed to provide all necessary gain, which allows the differential amplifier to operate with $R_{1}=R_{2}$ which is the simplest arrangement to maintain. In particular, the resistors can be matched in temperature response and other drift behavior quite accurately, which improves the precision of the differencing quite substantially.

The second main improvement made in the instrumentation amplifier is to reduce the common-mode gain of the amplifier. Instrumentation amplifiers are frequently constructed with very large gains ( 10000 X or more) and common-mode amplification in the first stage can easily overwhelm the ability of the differential amplifier to operate. To fix this problem, the input non-inverting amplifiers have their $R_{1}$ connected together rather than to ground as seen in Figure 3.11. The output of this configuration can be shown (Problem 3) to be

$$
V_{\text {out }}=\left(V_{b}-V_{a}\right)\left(1+\frac{2 R_{F}}{R_{G}}\right)
$$

This result shows another major convenience of the instrumentation amplifier configuration : the differential gain can be controlled by varying the value of a single resistor $\left(R_{G}\right)$.

While instrumentation amplifiers can be constructed from individual op-amps and resistors, they are more often obtained as integrated circuits. The IC contains the two input


Figure 3.11: The instrumentation amplifier.
stage op-amps and the entire differential stage. The integration of the differential amplifier's resistors into the IC greatly improves the stability and precision of the its operation. The user of the IC instrumentation amplifier need only provide the two $R_{F}$ and one $R_{G}$ resistors, as well as the power and input signals.

### 3.4.3 Current to Voltage and Voltage to Current

The circuit in Figure 3.12 looks like an inverting amplifier with (negative) infinite gain: $R_{1}=0 \Omega$. Indeed, if one connects a voltage source to the input, the result will necessarily be a saturated output with a polarity opposite to the input. However, this circuit is intended for converting current signals into voltage signals; it is a transimpedance amplifier. We can easily see this by observing the effect of a current signal $I_{i n}$ applied to the inverting input.

$$
\begin{aligned}
I_{\text {in }} & =I_{2} \\
& =-\frac{V_{\text {out }}}{R_{2}} \\
V_{\text {out }} & =-I_{\text {in }} R_{2}
\end{aligned}
$$

Thus, the magnitude of the output voltage is proportional to the input current and the resistor $R_{2}$. This may not seem like a major accomplishment - a simple resistor is able to convert a current into a voltage. Notice, however, the difference in input and output impedance for a transimpedance amplifier compared to a simple resistor. The input impedance of the transimpedance amplifier is zero, which the desired behavior for a current sink, while the input impedance of a resistor is $R$. The output impedance of the amplifier is also zero, while


Figure 3.12: A transimpedance or current-to-voltage amplifier.
the output impedance of a resistor to ground is $R$. Thus, the transimpedance amplifier allows the choice of $R$ (and thus the gain) to be independent of the input and output impedances, both of which are optimal for a current-in/voltage-out device.

Besides converting currents into voltages, we can use op-amps to convert voltages into currents. In Section 3.2.3, we studied the meter amplifier which is a voltage to current converter where the leads of the meter are connected between the output of the op-amp and its inverting input. Often, however, we wish to supply a current which will be returned through ground rather than having both leads connected to the op-amp. The circuit in Figure 3.13 is able to provide such a voltage to current conversion.

To ease the analysis of the voltage-to-current conversion circuit, Figure 3.13 includes labels for the currents flowing through each of the resistors including the load resistor $R_{L}-$ clearly the most interesting current is the current through the load! This current can be determined using Kirchhoff's current sum rule at the positive input of the op-amp. Taking into account the positive current directions as defined by the arrows, Kirchhoff's rules require:

$$
I_{c}=I_{d}+I_{L}
$$

It is simple to write down the current $I_{d}$ in terms of $V_{+}$.

$$
I_{d}=\frac{V_{+}}{R_{1}}
$$

We will make the assumption of the virtual short-circuit for this circuit, which tells us that $V_{+}=V_{-}$. We can use this fact to determine a relationship between $I_{c}$ and $I_{b}$

$$
I_{c}=\frac{V_{\text {out }}-V_{+}}{R_{2}}
$$



Figure 3.13: A voltage-to-current amplifier.

$$
=\frac{V_{\text {out }}-V_{-}}{R_{2}}=-I_{b}
$$

Since the inverting input of the op-amp draws no current, we can observe that

$$
\begin{aligned}
I_{b} & =I_{a} \\
I_{c} & =-\frac{V_{i n}-V_{+}}{R_{1}}
\end{aligned}
$$

With these results, we can simplify the current equation.

$$
\begin{aligned}
I_{c} & =I_{d}+I_{L} \\
\frac{V_{+}}{R_{1}}-\frac{V_{i n}}{R_{1}} & =\frac{V_{+}}{R_{1}}+I_{L} \\
I_{L} & =-\frac{V_{i n}}{R_{1}}
\end{aligned}
$$

Thus, the current through the load can be set by the input voltage, provided that the op-amp is not driven into saturation trying to supply sufficient voltage and current. This is not the only solution, of course; by using two op-amps, a current source can be created which does not exhibit the sign inversion present in this simpler circuit.


Figure 3.14: An op-amp based integrator circuit with reset.

### 3.4.4 Integrators and Differentiators

As we have seen, the inverting amplifier is a flexible and interesting circuit, the behavior of which can be modified in many ways by additional resistors. The behavior can be extended still further by replacing one or the other resistor in the inverting amplifier by a capacitor. Depending on which resistor is replaced, the resulting circuit operates either as a integrator or differentiator of an input voltage signal.

Let us begin with the integrator. In this circuit, we replace the feedback resistor $R_{2}$ with a capacitor as shown in Figure 3.14. For practical reasons, the circuit also includes a switch across the capacitor which will allow the capacitor to be discharged, resetting the integrator. To analyze the response of the system, consider the situation where the capacitor is initially discharged and a voltage signal $V_{i n}(t)$ is applied to the input. The input resistor will convert the voltage signal into a current $I(t)$.

$$
I(t)=\frac{V_{i n}(t)-V_{-}(t)}{R}
$$

This current will charge the capacitor, according to its "Ohm's Law".

$$
\begin{aligned}
V_{-}(t)-V_{\text {out }}(t) & =\frac{1}{C} \int_{0}^{t} I(t) d t \\
& =\frac{1}{R C} \int_{0}^{t} V_{\text {in }}(t)-V_{-}(t) d t
\end{aligned}
$$

The only problem is that we don't seem to know $V_{-}(t)$. However, we can actually apply the principle of the virtual ground even in this case - as the $\Delta V$ on the capacitor increases, the op-amp gain will push down $V_{\text {out }}$ to compensate to keep $V_{-}(t)=0$. This will continue until


Figure 3.15: Circuit designs for a simple (a) and realistic (b) differentiator using an op-amp.
the op-amp's output saturates. Thus, the output of the circuit becomes the negative integral of the input voltage signal.

$$
V_{\text {out }}(t)=-\frac{1}{R C} \int_{0}^{t} V_{\text {in }}(t) d t
$$

If the input resistor is removed, the integrator is converted into a transimpedance integrator operating directly on the input current.

For a realistic integrator, it is generally very important to use an op-amp with very small input bias currents or to balance out the bias current using an adjustable current source connected to the inverting input. Otherwise, the integration of the bias current may dominate the signal, particularly for long integration times.

If we instead replace the input resistor $\left(R_{1}\right)$ with a capacitor and keep the feedback resistor, the result is the differentiator shown in Figure 3.15a. The behavior of the differentiator is easy to demonstrate, as the principle of the virtual ground clearly applies here.

$$
\begin{aligned}
I_{C}(t) & =C \frac{d\left(V_{\text {in }}(t)-V_{-}(t)\right)}{d t} \\
& =C \frac{d V_{\text {in }}(t)}{d t}=I_{R} \\
\frac{V_{-}-V_{\text {out }}(t)}{R}=-\frac{V_{\text {out }}(t)}{R} & =C \frac{d V_{\text {in }}(t)}{d t} \\
V_{\text {out }}(t) & =-R C \frac{d V_{\text {in }}(t)}{d t}
\end{aligned}
$$

This differentiator design, however, suffers from a significant flaw. For high frequency signals, the impedance of the capacitor will go to zero $\left(Z_{C}(\omega) \rightarrow 0\right)$. This is equivalent to taking $R_{1} \rightarrow 0$ in the conventional inverting amplifier, which means that a differentiator built as in Figure 3.15a will suffer from a very large amplification of high frequency noise. This noise will dominate any signal from the differentiator. One way to attack this problem is shown in Figure 3.15b. The combination of $R_{1}$ and $C$ acts as a low-pass filter with a characteristic frequency $R_{1} C$. However, this filtering behavior may distort the differential


Figure 3.16: An op-amp-based comparator.
of $V_{i n}(t)$. Because of these significant non-idealities, the differentiator is rarely used in practice.

### 3.4.5 Comparators

All of the uses of the op-amp to this point in the chapter have involved negative feedback and the principle of the virtual short circuit has applied. There is one use where this principle certainly does not apply: the comparator as seen in Figure 3.16. This is a strikingly simple circuit where we cannot assume $V_{-}=V_{+}$. In fact, whenever $V_{i n}<V_{-}$, the output of the circuit will go into negative saturation while when $V_{i n}>V_{-}$, the output will be in positive saturation. The comparator thus performs a threshold function, which is often useful to trigger some action for coupling analog circuits to the digital circuits we will discuss in later chapters.

An op-amp is actually not the best choice for this circuit. In particular, general-purpose op-amps have significant delays in coming out of saturation, so they tend to be quite slow as comparators. In addition, most digital logic systems (as we'll see in detail later) would prefer to receive either $+V$ or ground, rather than the negative and positive rails of the opamp power supply. As a result, dedicated comparator chips exist which avoid full saturation and provide an open-collector output. An example circuit using such a comparator is shown in Figure 3.17. An open-collector output is one which will sink current but not supply any. This means that the output of the comparator is

$$
\begin{cases}0 V & V_{\text {in }}<V_{-} \\ 0 A\left(Z_{o} \rightarrow \infty\right) & V_{\text {in }}>V_{-}\end{cases}
$$

As a result, the final output voltage will be zero for inputs below threshold and $V_{c c}$ for inputs above threshold. The value of $V_{c c}$ can be tuned for the particular flavor of digital logic in use.


Figure 3.17: An open-collector output comparator allowing connection to any digital logic family.

## Problems

1. Consider the inverting amplifier with capacitively coupled inputs below. Show that the amplification of high frequency signal ( $\omega \gg 10 \mathrm{kHz}$ ) is the same if the signal is applied to either Point A or Point B. Next, find the ratio between the response of the circuit due to the offset voltage $V_{o s}$ if the input signal is applied to the two different positions. Does this effect depend on temperature? This demonstrates one way to manage offset voltages in high-gain designs.

2. Consider the schematic below of a sensor and an amplifier. Determine the voltages at $V_{\text {in }}$ and $V_{\text {out }}$ in terms of $V_{s}$ and compare to the expected gain of the standalone amplifier. Can you explain any observed difference? How could the difference be decreased?

3. Show that for the circuit shown in Figure 3.11

$$
V_{\text {out }}=\left(V_{b}-V_{a}\right)\left(1+\frac{2 R_{F}}{R_{G}}\right)
$$

Hint: Since the right portion of the circuit is simply a differential amplifier (with the behavior already derived in Section 3.4.1), you need only focus on the left portion of the circuit and determine the difference between the inputs to the differential amplifier stage.

## Chapter 4

## Sensors

For electronics to be useful for an experimenter, the physical (or chemical) observables of an experiment must be converted into electrical signals. This is the job of the sensor - a wide class of devices which convert sound, temperature, pressure, light, acceleration, and more into electrical currents or voltages. Once the conversion is made, we can apply all the amplification and processing techniques we have learned using op-amps to tailor the signal to our needs.

This chapter will cover both the basic principles of sensors and of the design of measurement circuits. Measurement precision can often be improved either by the choice of more precise (and more expensive) components and instruments or by careful design of the measurement to limit the impact of noise on the measurement.

### 4.1 Basic Sensor Principles

In general, sensors can be modeled as combinations of resistors, capacitors, inductors, voltage sources, and current sources. A given type of sensor will generally have a dominant characteristic, such as being primarily resistive in nature with a resistance proportional to the quantity to be measured. The appropriate measurement technique will depend on the nature of the sensor. Generally, the final goal is to produce a voltage signal which can be easily converted to digital form for acquisition and analysis with a computer.

Besides the nature of the sensor, an additional important factor is the environment of sensor and its connection to a measurement apparatus. For example, cryogenic experiments frequently involve measurements where the sensor is kept within fractions of a degree of absolute zero while the electronics is kept at room temperature. Such configurations necessarily lead to long lead wires between the sensors and the amplifiers. This can lead to significant noise pickup on the sensor leads. All wires can act as antennae receiving ambient electromagnetic radiation. This antenna-like behavior results in unwanted additional signals on the lead wires which is one of the major sources of noise in an experimental setup.

The induced noise is most often in the form of a weak, time-varying current source. As a result, the largest noise impact is observed for amplifiers with very high input impedance.


Figure 4.1: Reference-line technique for subtracting common-mode noise using an instrumentation amplifier.

In this case, a small additional current can become a large voltage signal through Ohm's Law. Unfortunately, amplifiers with large input impedance are needed to amplify weak (high output impedance) sensors. These amplifiers must either be installed very near the sensors (as "pre-amplifiers") or shielding should be used to reduce the noise pickup.

To reduce the induced noise in wires and cables, it is often necessary to surround the signal wire with a metal sheath (often a flexible braid) which is connected to ground. This grounded shield will absorb the external noise while the signal wire - a Faraday cage for a single wire. As an added benefit, the shield provides a constant potential environment for the signal and a constant characteristic impedance which improves signal integrity. Similarly, low noise amplifiers are usually enclosed in a metal box for shielding purposes. In all cases, it is good design to minimize the length of wires used to assemble a device and to avoid loops or similar structures which can act as even more effective antennae than simple straight wires.

### 4.1.1 Voltage and Current Sensors

The simplest sensors to use are those which provide voltage signals directly. For such a sensor, the main concerns are the output impedance of the sensor and any noise which may be picked up along the way. A non-inverting amplifier (such as a unity-gain buffer amplifier) can improve the impedance characteristics of a voltage sensor, and such initial "preamplifiers" are often found at the beginning of a signal processing chain. Noise management for voltage sensors often includes both passive techniques such as using well-shielded cables as well as using differential measurement techniques to remove the common-mode noise.

Inductive noise as well as ground noise and other effects at the sensor can be subtracted using a reference line technique as seen in Figure 4.1. Rather than using a single voltage sig-


Figure 4.2: Direct constant-current (a) and constant-voltage (b) techniques for resistivesensor measurements.
nal, the design uses two signals with similar characteristic impedances $\left(Z_{s i g} \approx Z_{r e f}\right)$. Most noise sources, including the inductive noise represented by $V_{\text {noise }}$, will induce commonmode swings in the two lines - both lines will rise or fall together as a result of the inductive noise. The reference line can then be used to subtract the common-mode noise by feeding the reference and measurement lines into an instrumentation amplifier.

Current-mode sensors tend to be much more robust against noise than voltage-mode sensors, particularly when low input-impedance amplifiers are used. Conventionally, $50 \Omega$ characteristic impedance inputs are often used which match well to a wide range of coaxial cables. Twisted pair cables often have characteristic impedances around $100 \Omega$. Sensors which produce current signals can use a simple resistor or a transimpedance amplifier (Section 3.4.3) to convert the current signal into a voltage. The transimpedance amplifier has the advantage of an ideal zero-impedance for the current input, compared to the $Z_{i n}=R$ for the simple resistor. The main disadvantage of the simple transimpedance amplifier is that the op-amp input is directly connected to the outside world and may be damaged by excessive voltages or currents - the addition of a series resistor raises the input impedance but provides protection.

### 4.1.2 Resistive Sensors

Many sensors are not pure current or voltage sources, but rather behave as resistors with a resistance proportional to the quantity to be measured. A resistive sensor will generally depend on an Ohm's law-based measurement. A direct application of Ohm's law would imply applying a known current and measuring a voltage as in Figure 4.2a or applying a known voltage and measuring a current as in Figure 4.2b. In both figures, the smaller resistors represent parasitic resistances from the cabling and contacts. These resistances will induce errors in the measurement: the actual resistance being measured is $R+R_{\text {wire }}+R_{\text {wire }}^{\prime}$. If $R \gg R_{\text {wire }}$ then the effect of the wire resistance can generally be neglected, otherwise a technique such as the four-point resistance measurement described in Chapter 2 may be used.


Figure 4.3: Wheatstone bridge and instrumentation amplifier combination for precise measurement of small changes in resistance.

If the response of the sensor is linear, we can write the variation of resistance with the quantity under measurement $(x)$ as

$$
R(x)=R_{\circ}+\Delta_{R}(x)=R_{\circ}+K_{R}\left(\frac{x-x_{\min }}{x_{\max }-x_{\min }}\right)
$$

where $x_{\text {max }}$ and $x_{\text {min }}$ are the largest and smallest values of the parameter which will be observed. The resistances change coefficient $K_{R}$ thus has units of resistance representing the change of resistance over the full parameter range. Often, the change in resistance for a resistive sensor is quite small over the full dynamic range of the measurement compared to the total resistance, so $C_{R} \ll R_{\circ}$. Thus, we are interested in a small change in resistance on top of a large constant resistance. It can be difficult to make this measurement accurately as increasing precision in an amplifier limits the dynamic range of the amplifier - we must avoid amplifier saturation.

The natural alternative to a direct resistance measurement is a differential measurement which can bring the power of the instrumentation amplifier into play. One particularly powerful arrangement is the Wheatstone bridge configuration shown in Figure 4.3. In this configuration, a voltage is applied across a pair of voltage dividers in parallel and the measurement is made between the two inner nodes of the bridge.

$$
\begin{aligned}
V_{1} & =V_{c c} \frac{R_{2}}{R_{\text {meas }}+R_{2}} \\
V_{2} & =V_{c c} \frac{R_{2}}{R_{1}+R_{2}}
\end{aligned}
$$

$$
V_{2}-V_{1}=V_{c c}\left(\frac{R_{2}}{R_{1}+R_{2}}-\frac{R_{2}}{R_{\text {meas }}+R_{2}}\right)
$$

Thus, if $R_{1}=R_{\text {meas }}$, the voltage difference will be close to zero. Any small deviations in $R_{\text {meas }}$ will cause changes around zero rather small changes on a large constant value, allowing the instrumentation amplifier and any following amplifiers to provide a large gain.

Substituting $R_{1}=R_{\circ}$ into the equation for the voltage difference, we find

$$
\begin{aligned}
V_{2}-V_{1} & =V_{c c}\left(\frac{R_{2}}{R_{\circ}+R_{2}}-\frac{R_{2}}{R_{\circ}+\Delta_{R}(x)+R_{2}}\right) \\
& =V_{c c}\left(\frac{1}{R_{\circ} / R_{2}+1}-\frac{1}{R_{\circ} / R_{2}+\Delta_{R}(x) / R_{2}+1}\right)
\end{aligned}
$$

If we consider the case when $\Delta_{R}(x) \ll R_{2}$, we can Taylor-expand this expression using $\frac{1}{a+x}=\frac{1}{a}-\frac{x}{a^{2}}+O\left(x^{2}\right)$.

$$
\begin{aligned}
\frac{1}{R_{\circ} / R_{2}+1+\Delta_{R}(x) / R_{2}} & \approx \frac{1}{R_{\circ} / R_{2}+1}-\frac{\Delta_{R}(x) / R_{2}}{\left(R_{\circ} / R_{2}+1\right)^{2}} \\
V_{2}-V_{1} & \approx V_{c c}\left(\frac{\Delta_{R}(x) / R_{2}}{\left(R_{\circ} / R_{2}+1\right)^{2}}\right)
\end{aligned}
$$

Thus, the change in resistance is directly proportional to the observed voltage difference.

### 4.1.3 Capacitive and Inductive Sensors

Some sensors, including many position sensors, behave as variable capacitors or inductors. To measure the change in capacitance or inductance, a AC excitation is generally needed as the DC voltage across a capacitor or inductor is independent of the device's value. One straightforward way to convert a capacitance or inductance change into a voltage is to use the device as part of a low-pass or high-pass filter and then determine the average voltage using a diode and low-pass stage. A schematic for using this technique to measure a changing capacitance is shown in Figure 4.4.

The measurement must be driven by a sinusoidal voltage with a characteristic frequency $\omega_{i n}$. If we choose $\omega_{i n}$ and $R_{1}$ to put the input filter below its roll-off (in the $-20 \mathrm{~dB} /$ decade regime) for the expected capacitance ranges, we find that

$$
\frac{V_{\text {filt }}(t)}{V_{\text {in }}(t)} \approx \omega_{i n} R_{1} C_{\text {meas }}
$$

The amplitude ratio of $V_{\text {filt }}$ over $V_{i n}$ is thus directly proportional to the capacitance. The role of the buffer in the circuit is to provide a impedance conversion of the filtered signal to drive the final "detector" diode and low-pass filter. The role of the diode and the final lowpass filter is to convert the oscillating signal with an average voltage of zero into a constant voltage close to $V_{\text {filt }}$. The utility of this design can be limited by coupling of high-frequency noise, so it is sometimes necessary to change the input filter to a low-pass type. This yields an inverse proportionality between $V_{\text {filt }}$ and the variable capacitance.


Figure 4.4: Measurement of a changing capacitance using a high-pass filter technique.

### 4.1.4 Semiconductor Device Physics

Many interesting sensors are based on semiconductor materials, so a review of semiconductor device physics is relevant before discussion photodiodes and other semiconductor sensors. We will not make a complete derivation of the relevant physics here - the student is recommended to consult a full semiconductor device textbook for complete details if needed.

From a condensed-matter viewpoint, a semiconductor is a material where at zero temperature the highest-energy occupied band of states (the valence band) is completely full and next higher band (the conduction band) is empty. At zero degrees, the semiconductor will not be able to conduct current. This contrasts with a metal, where the highest-energy occupied band is not full and conduction can occur. An insulator has the same band structure characteristics as semiconductor - the distinction is that the energy gap between the bands is small (a few times $k_{B} \times 300 \mathrm{~K}$ ) for a semiconductor and much larger for a good insulator. At temperatures above absolute zero, thermodynamic effects will cause the generation of electron-hole pairs at a rate dependent on temperature and the band-gap. Thus, a pure semiconductor at finite temperature will be a poor conductor of current, but not an insulator. Notice that whenever a electron is promoted from the valence band to the conduction band, a gap (or hole) is left in the valence band. This hole functions as an effective charge carrier with positive charge and can move through the material.

The characteristics of a semiconductor material can be modified by adding small quantities of impurities, often called dopants. These impurities tend to have states which lie in the valence-conduction gap. The states may lie near the valence band, where they tend to be empty states at $T=0$, or they may be near the conduction band where they tend to be full at $T=0$. Notice that the concentration of the dopants is low enough that these states do not form bands - they remain as isolated states. At finite temperature, dopants with states near the conduction band will lose their electrons into the conduction band, providing an excess of free electrons. Thus, a material doped with donor impurities is called an "n-type" material, as it has an excess of negative charge carriers. On the other hand, acceptor dopants at finite temperature will trap electrons from the valence band, leaving extra holes in the valence band. Materials doped with acceptor impurities are thus called "p-type" materials.


Figure 4.5: Band-diagram structure for a pure-semiconductor (a), doped semiconductor at $T=0 \mathrm{~K}(\mathrm{~b})$, doped semiconductor at finite temperature.

When n-type and p-type materials are connected together, the result is a pn junction. Along the interface between the two materials, the excess holes from the p-type material and the excess electrons from the n-type material will combine and as charge carriers diffuse to the junction, the process continues. However, the original donor or acceptor atoms are fixed in the lattice and unable to move so this process creates a population of positivelycharged ions in the n-type material and negatively-charged ions in the p-type material as seen in Figure 4.6. These ions set up an electric field which opposes any further motion of holes from the p-type or electrons from the n-type material and establishes an inherent potential barrier or contact potential in the device. The region over which the electric field extends has no free charge carriers and is generally referred to as the depletion region. The width of the depletion region along the axis between the electrodes is called the depletion width and is a critical parameter in junction design.

Because of its built-in electric field, a pn-junction is able to function as a diode. If the p -electrode is raised to a higher voltage than the n -electrode, the externally-applied electric field will cancel part of the ion-induced field, allowing a current to flow. However, if the n -electrode is raised to the higher potential, the depletion region will be widened and no current will flow. Thus, the p-type end of the diode is the anode and the n-type is the cathode.

### 4.2 Selected Sensors

Now that we have discussed the general principles of sensors and measurement, we can focus on a specific subset of sensors.

### 4.2.1 Temperature Sensors

Temperature sensors are important both for measurement and for control and monitoring of experiments. For cryogenic measurements, for example, an accurate measurement of the


Figure 4.6: A pn-junction
temperature is critical to understand superconducting transitions, ferromagnetic behavior, and many other condensed matter parameters. There are several different categories of sensors which can be used.

## Thermistors

The least expensive but least accurate solution for temperature measurement is the thermistor. Recall from Section 2.6.1 that real resistors have temperature-dependent behavior which can be taken as linear for small variations in temperature. Regular resistors are designed to minimize this effect, but thermistors are designed to make these effects large.

$$
\frac{\Delta R}{R}=K \Delta T
$$

Depending on the material that a resistor is built from, the temperature coefficient may be positive or negative. In general, semiconductor-like resistors have negative temperature coefficients (resistance falls with increasing temperature) while metallic resistors have positive temperature coefficients.

Thermistors in general are not tightly specified; one cannot simply read a resistance and infer an absolute temperature. Instead, a calibration procedure must be followed to establish the resistance at a known temperature and the temperature coefficient can be used to follow temperature changes around that known value. The differential resistance measurement required for a thermistor is very well matched to the bridge technique. Sometimes additional terms are included to the $\Delta R(T) / R$ relationship to extend the performance of the device outside the linear region.

## Special Topic: Light Emitting Diodes

Light emitting diodes (LEDs) are an important subcategory of the diode family. An LED is essentially a conventional diode built of a material with a large band-gap (on the order of $2-3 \mathrm{eV}$ ). When a diode is operated in forward conduction, electrons and holes are injected at the cathode and anode respectively and combine in the depletion region with a release of energy. In the materials used to build LEDs, the transition involves only a change in energy (not momentum) which means that a photon with energy equal to the band-gap is produced. In materials like silicon, the transition requires a momentum change of the electron and energy which is released is generally absorbed by the silicon lattice as a phonon rather producing a photon.
LEDs are extremely efficient at converting electrical energy into light and are used both as indicators and increasingly as general-purpose light sources. Besides providing the power and indicator lights for many devices, LEDs provide the back-light for many laptop and telephone screen and are used in many new stoplights. To use an LED, one must provide a current source of $\sim 10 \mathrm{~mA}$. Generally, one uses a voltage source, however, so a current-limiting resistor is required as seen below. To determine the necessary resistor, we assume that the LED provides a constant voltage drop - the magnitude of the drop depends on the band-gap of the material and thus the wavelength of the light produced. With this voltage drop, we choose the resistor to limit the current to the desired level.

$$
\begin{aligned}
\Delta V & =I R \\
R & =\frac{V_{c c}-V_{f b}}{I}
\end{aligned}
$$




Figure 4.7: Temperature measurement using iron/constantan $(\mathrm{CuNi})$ thermocouples with an ice-bath as a reference temperature.

Since a thermistor is a resistor, injection of a current to measure a voltage difference will result in power dissipation in the thermistor. This implies a self-heating of the thermistor independent of the temperature it is measuring. As a result, the thermistor should be thermally well-connected to the sample it is measuring, the current through the thermistor should be as small as possible, and the sample heat capacity should be large enough to absorb the heat load from the thermistor without a large increase in temperature.

## Thermocouples

An alternative physical principle which can be used for temperature dependence is the Seebeck effect. When a temperature gradient is applied along any metal, a potential difference develops which is due to the differences in carrier mobility as a function of temperature. To measure this voltage difference, however, one must use a probe which will necessarily experience the same temperature difference. If the probe is made of the same metal, the potential differences will cancel. Fortunately, the magnitude of the Seebeck effect is different for different metals, so a junction of different materials can be used the measurable voltage difference will be the difference in Seebeck effects between the two metals. This is the basis for the thermocouple - two dissimilar metals connected at a single point. Common metal pairs for thermocouple construction include Nickel-Chromium/Nickel-Aluminum and Iron/Copper-Nickel. In the linear approximation,

$$
\Delta V=\alpha\left(T-T_{0}\right)
$$

where $\alpha=40 \mu \mathrm{~V} / \mathrm{K}$ for $\mathrm{NiCr} / \mathrm{NiAl}$ and $\alpha=51 \mu \mathrm{~V} / \mathrm{K}$ for $\mathrm{Fe} / \mathrm{CuNi}$.
In a practical thermocouple-based system, we must also worry about the junction between the thermocouple and the measurement leads which connect the thermocouple to the amplifier. These will effectively provide two additional thermocouples at approximately


Figure 4.8: Example circuit showing the use of an AD590 semiconductor temperature sensor with an output scale of $1 \mathrm{mV}=1 \mathrm{~K}$.
room temperature. In such an arrangement it is very hard to calibrate the sensor and perform an accurate measurement. Instead, usually a second thermocouple is used connected so the that the two thermocouples are connected by a single type of metal. The measurement is thus the difference in temperature between the two thermocouples. For precision measurements, the second thermocouple is placed into a precision temperature reservoir such as an ice bath. A schematic of a full system based on thermocouples is shown in Figure 4.7. Since the thermocouple is a welded combination of pure metals, it can be used to very high temperatures. Thermocouples are widely used to measure temperatures in metallurgy and flame temperatures including safety systems for pilot lights. For these purposes, an accurate low-temperature reference is often irrelevant.

## Current-mode temperature sensors

Thermocouples are effective for a wide range of temperatures, but the requirement for a reference temperature point is inconvenient and occasionally problematic. One alternative for temperatures in the range $-50^{\circ} \mathrm{C}<T<200^{\circ} \mathrm{C}$ is a semiconductor-based measurement. For example, transistors and diode both exhibit temperature-dependence in their I-V curves, which can be used to measure temperature. The use of a semiconductor sensor allows the integration of additional components to the sensor circuit as part of the IC at very low cost. These components act as a local amplifier/signal processor which gives the full device a simple interface.

An example sensor of this type is the Analog Devices AD590 sensor, which is designed and tuned to provide a current signal which is exactly proportional to the sensor's absolute temperature, when at least 4 V of bias is applied across the two terminals of the device as
shown in Figure 4.8.

$$
I(T)=\frac{1 \mu \mathrm{~A}}{1 K} T
$$

Thus, the current in microamps directly gives the absolute temperature in Kelvin. The current signal can be converted to a voltage signal either by a simple resistor as in Figure 4.8 or by a transimpedance amplifier configuration. These sensors are particularly convenient when long wires are needed as the current-mode signal is relatively insensitive to noise pickup and the additional resistance of the long leads does not affect the current signal as long as sufficient voltage is available at the device.

### 4.2.2 Photosensors

Photosensors are important both for direct measurements of light and also in control and communications applications. Semiconductors are well-matched to the task of photon detection. Electrons in the valence band of a semiconductor can be promoted to the conduction band by absorbing a photon, forming an electron-hole pair.

## Photoresistors

Photoresistors are the simplest type of semiconductor photodetector : a single piece of intrinsic (undoped) semiconductor. The electron-hole pairs produced as light is absorbed by the photoresistor will increase the conductance of the semiconductor, reducing its resistance. Photoresistors are commonly built from semiconductors with large band-gaps to minimize the number of thermally-produced electron hole pairs. This increases the relative sensitivity of the device to light, since the ratio of photo-produced electron-hole pairs to thermal pairs is increased. Photoresistors are often used in control circuits, as their sensitivity is relatively low and their response to changing light levels is relatively slow.

## Photodiodes

The workhorse of silicon photosensors is the photodiode. Photodiodes are operated in reverse-biased configuration as shown in Figure 4.9. When the diode is reverse-biased, the current flow is very small as for any diode. When electron-hole pairs are photoproduced in the depletion region, the high electric field in the depletion region rapidly separates the electron and hole pair. The photo-produced carriers provide most of the current in the device, so a photodiode operates as a current source and the appropriate amplifier is the transimpedance amplifier as shown in Figure 4.9. The noise level in photodiodes is quite low since the number of free thermal carriers is small and high internal field of diode makes it quite fast - photodiodes used in communications applications are able to resolve pulses with resolution of 50 ps .


Figure 4.9: Light amplitude measurement with a photodiode and transimpedance amplifier.

### 4.3 Advanced Measurement Techniques

For small signals hidden in large constant backgrounds and large quantities of noise, more advanced techniques may be needed. One of the most powerful techniques is the lock-in amplifier.

### 4.3.1 The Lock-In Amplifier

In the lock-in technique, the amplifier provides a reference signal which is to be connected to the apparatus and then multiplies the reference signal with the input from the sensor. The output of the sensor must be proportional to the reference, with a possible phase shift. The multiplication could be provided by a diode using the part of the diode characteristic curve where $I \propto V^{2}$ or could be performed using a digital technique. The term "mixer" is used to generically designate device which performs the multiplication. The input voltage to the mixer is the superposition of the reference signal, the measured signal, and noise. We can write these as

$$
\begin{aligned}
V_{\text {ref }}(t) & =A_{r} \cos \left(\omega_{r} t\right) \\
V_{\text {sense }}(t) & =A_{s} \cos \left(\omega_{r} t+\delta\right) \\
V_{\text {noise }}(t) & =\sum_{i} A_{i} \cos \left(\omega_{i} t\right)
\end{aligned}
$$

The noise is represented as a Fourier series of noise components, each with its own frequency and amplitude. The amplitude of any particular component can be considered to be small but the number of terms in the noise will be large.

Given this, we can write the output voltage of the mixer as a function of time.

$$
\begin{aligned}
V_{\text {mixer }}(t)= & 2 K\left[V_{\text {ref }}(t) V_{\text {sense }}(t)+V_{\text {ref }} V_{\text {noise }}(t)+V_{\text {sense }}(t) V_{\text {noise }}(t)\right] \\
= & 2 K\left[A_{r} A_{s} \cos \left(\omega_{r} t\right) \cos \left(\omega_{r} t+\delta\right)+\right. \\
& \left.\sum_{i} A_{r} A_{i} \cos \left(\omega_{r} t\right) \cos \left(\omega_{i} t\right)+\sum_{i} A_{s} A_{i} \cos \left(\omega_{r} t+\delta\right) \cos \left(\omega_{i} t\right)\right]
\end{aligned}
$$

We can simplify this equation using a trigonometric identity.

$$
\cos (A) \cos (B)=\frac{1}{2} \cos (A+B)+\frac{1}{2} \cos (A-B)
$$

When we apply this to the mixer output voltage equation, we obtain a large number of terms involving the sums and differences between frequencies.

$$
\begin{aligned}
\frac{1}{K} V_{\text {mixer }}(t)= & A_{r} A_{s} \cos \left(2 \omega_{r} t+\delta\right)+A_{r} A_{s} \cos (-\delta)+ \\
& \sum_{i}\left[A_{r} A_{i} \cos \left(\left[\omega_{r}+\omega_{i}\right] t\right)+A_{r} A_{i} \cos \left(\left[\omega_{r}-\omega_{i}\right] t\right)\right]+ \\
& \sum_{i}\left[A_{s} A_{i} \cos \left(\left[\omega_{r}+\omega_{i}\right] t+\delta\right)+A_{s} A_{i} \cos \left(\left[\omega_{r}-\omega_{i}\right] t+\delta\right)\right]
\end{aligned}
$$

For the lock-in amplifier, the mixer is followed by a low-pass filter with a very low characteristic frequency. As a result, all the terms involving $\omega_{r}$ and $\omega_{i}$ drop out except for any component which is exactly coherent with the reference frequency. If the reference frequency is well-chosen to be far from common peaks in the noise spectrum (such the 60 Hz line noise), then we can neglect the noise contributions in the final current relationship.

$$
V_{\text {out }}=K A_{r} A_{s} \cos (\delta)
$$

The resulting measurement is thus dependent on the phase between the sensor and reference signals as well as the magnitude of the sensor signal. This is inconvenient, so most lock-in instruments provide a second internal channel where the reference frequency is phase-shifted by $\frac{\pi}{2}$. The result is two measurements which can be combined to yield both the amplitude and phase of the sensor signal.

$$
\begin{aligned}
V_{a} & =K A_{r} A_{s} \cos (\delta) \\
V_{b} & =K A_{r} A_{s} \cos \left(\delta+\frac{\pi}{2}\right)=-K A_{r} A_{s} \sin (\delta) \\
\sqrt{V_{a}^{2}+V_{b}^{2}} & =K A_{r} A_{s} \sqrt{\cos ^{2} \delta+\sin ^{2} \delta}=K A_{r} A_{s} \\
-\frac{V_{b}}{V_{a}} & =\frac{K A_{r} A_{s} \sin \delta}{K A_{r} A_{s} \cos \delta}=\tan \delta
\end{aligned}
$$

## Chapter 5

## Transistors

Transistors are the underlying building blocks for operational amplifiers as well as the digital circuits discussed later. However, their detailed behavior is relatively complex, compared to either operational amplifiers or digital circuits. In particular, questions of biasing and linearity must be addressed for discrete transistor circuits. As a result, experimentalists are recommended to begin with op-amps and other "packaged solutions" and use transistors only when necessary.

However, it is worthwhile to have a basic introduction to the concepts of transistors to help in the understanding of misbehavior in other circuits. Also, transistors are also particularly useful for control applications where large currents or voltages may be needed. In these cases, the transistors can be "managed" by an op-amp or logic chip, making the use of the transistor relatively simple. Use of discrete transistors is generally mandatory for currents above 100 mA , and many transistors are available for such currents in packages which allow for efficient heat-dissipation.

This chapter serves as a basic introduction to transistors and covers operations where the detailed behavior of the transistors is not crucial. More-advanced design work should refer to a more-detailed text.

### 5.1 Basic Transistor Operation

Transistors are three-terminal devices. Two of the terminals define a primary current-flow path which can be controlled in some manner by the third terminal. At a basic level, the control can be considered to be either current-based or voltage-based. This distinction divides transistors into two groups: the bipolar junction transistors (BJTs) where the control is current-based and the field-effect transistors (FETs) where the control is voltage-based. Historically, BJTs were the first class of transistors developed, while FET transistors were developed later but have come into very wide use due to their lower power consumption in digital logic.


Figure 5.1: Schematics and current flow labels for npn (left) and pnp (right) bipolar junction transistors.

### 5.1.1 Bipolar Junction Transistors

Bipolar junction transistors are structurally quite similar to a "diode-and-a-half". A diode contains a junction between a p-type region and an n-type region, as described in Chapter 4 with the p region representing the anode and the n region the cathode. As a result, current is allowed to flow from n to p , but not in reverse. In a bipolar transistor, an additional n or p layer is added to create either an "npn" or "pnp" transistor. Each of the regions is connected to a lead and has a particular role. The central region (which may not be the middle lead of the package!) is called the gate and it is responsible for the control. The other two leads are called the emitter and the collector.

The schematics for the two types of bipolar junction transistors are shown in Fig. 5.1. The arrow indicates the normal flow of current during operation. Thus, for a npn transistor the main current flow is from the collector to the emitter while for a pnp it from the emitter to the collect. A careful consideration indicates that the current flow between the base and the emitter is forward-biased in both cases ( p to n ) while the flow across the whole device contains the backward-biased collector-base junction. Nonetheless, current will flow between the collector and emitter when a smaller current is injected into the base-emitter system. The currents are proportional through a factor $\beta$ called the current gain.

$$
I_{C}=\beta I_{B}
$$

Typical values of $\beta$ range from 40 to 200, depending on the transistor technology. This simple proportionality suggests easy use of transistors for current application - just pick the right $\beta$ (like a resistance) and away you go! Unfortunately, $\beta$ is a difficult-to-control parameter which varies significantly from device-to-device.

However, a basic voltage amplifier can be constructed using a transistor and external resistors as shown in Fig. 5.2. To analyze this circuit, we must first consider that the diode junction formed by the base-emitter pair will have a forward bias voltage $V_{B E}$. Given this, we can determine the current through $R_{1}$, which is $I_{B}$ :

$$
\Delta V=I_{B} R_{1}
$$



Figure 5.2: Common-emitter npn amplifier

$$
\begin{aligned}
V_{i n}-V_{B E} & =I_{B} R_{1} \\
I_{B} & =\frac{V_{i n}-V_{B E}}{R_{1}}
\end{aligned}
$$

We can then apply the definition of the current-gain to determine the current through $R_{2}$, which is $I_{C}$, assuming zero current flows out the output.

$$
I_{C}=\beta I_{B}=\frac{\beta}{R_{1}}\left(V_{i n}-V_{B E}\right)
$$

The output voltage can thus be simply determined as

$$
\begin{aligned}
V_{c c}-V_{\text {out }} & =I_{C} R_{2} \\
& =\beta \frac{R_{2}}{R_{1}}\left(V_{i n}-V_{B E}\right) \\
V_{\text {out }} & =V_{c c}-\beta \frac{R_{2}}{R_{1}}\left(V_{i n}-V_{B E}\right)
\end{aligned}
$$

Thus, the circuit is roughly similar to an inverting amplifier with a gain $\beta \frac{R_{2}}{R_{1}}$, though the output signal is referenced to $V_{c c}$ rather than to ground. Such an amplifier is clearly able to handle only positive voltage signals, since the input voltage must be larger than the forward bias of the transistor. If the input voltage falls below the forward bias voltage, the current will fall to zero, clipping the waveform.

These various complexities in transistor operation are the original motivation for the creation of the op-amp, which combines approximately twenty transistors to produce its convenient characteristics.


Figure 5.3: Schematic symbols for n-channel and p-channel MOSFET transistors.

### 5.1.2 Field-Effect Transistors

The second major class of transistors behaves in a manner somewhat similar to a variable resistor controlled by a capacitor. In these transistors, the electric field from the control lead modifies the connection between the main current leads to change the conductance of the connection. The primary current leads in FET devices are called the source and drain. As in the case of BJTs, FET devices contain pn junctions which can block current flow. The application of an electric field to the control electrode, called the gate, causes the junction to be modified changing the conductance of the system. There are several types of FET in use for various purposes. We will consider the most commonly used type for digital electronics as an example - the enhancement-mode MOSFET.

In MOSFET devices, the source and drain electrodes connect to material of the same type ( p or n ), with a large region of the other type of material in between. Above the connecting region, but insulated from it by an oxide layer, is the gate electrode. The metal electrode followed by the oxide layer give the device its name: Metal-Oxide-Semiconductor FET. By applying a voltage to the electrode, electrons can be attracted to the connecting region or driven away. Thus, by applying a positive voltage, nominally p-type material can be temporarily changed into n-type material near the oxide layer. This provides a continuous channel with no reverse-biased junction between the source and drain and current flows. By modifying the voltage on the gate, the volume of material changed from p to n type can be adjusted, changing the conductance of the device.

MOSFETs can be operated in one of two modes, depending on the voltage between the drain and the source. In the ohmic mode, the voltage between the drain and source is less than the voltage between the gate and source. In this case, the current from the drain to the source can be modeled as

$$
I_{D S}=\xi\left[\left(V_{G S}-V_{t h}\right) V_{D S}-\frac{V_{D S}^{2}}{2}\right]
$$

where $\xi$ depends on the charge carrier density and geometry of the device and $V_{t h}$ is the gate threshold voltage, typically 1 V or less. In this mode, the current is proportional to the
gate-source voltage (for a constant drain-source voltage) which is why this mode is called ohmic. When the drain-source voltage is equal or larger than the gate-voltage, the device is in the saturation regime with a drain-source current which can be modeled as

$$
I_{D S}=\frac{\xi}{2}\left(V_{G S}-V_{t h}\right)^{2}
$$

This mode is called the saturated or active mode of the transistor, with the current proportional to the square of the gate-source voltage. Amplifiers with MOSFETs often operate in the saturated regime and apply to small signals around a biasing voltage where the drainsource current can be taken as approximately linear for small changes in $V_{C S}$.

The major advantage of MOSFETs arises from the zero current flow from the gate and the rest of the circuit. This provides a very large input impedance, which is often quite useful. In fact, this characteristic is crucial to modern digital electronics, which allows far lower power dissipation that would be the case with BJT-based circuitry. However, linear design with MOSFETs can be complex.

### 5.2 Use of Transistors

The amplifier circuit shown for the BJT indicates the complexity of arranging transistor circuits which are simple to integrate into measurements. However, transistors can be simply used in two cases: switching and power-boosting. In the case of switching, the transistor is only required to be in one of two states - fully conducting or off. This is relatively simple to arrange and allows the control of very high-current loads quite simply. In the case of power-boosting, a transistor can be added to an op-amp circuit and the natural feedback of the circuit used to keep the transistor behavior in line.

### 5.2.1 Switching

For switching applications, the n-channel MOSFET is an excellent device choice if the device which needs to be supplied is does not need to be grounded. An example circuit of this circuit is shown in Fig. 5.4. The basic operation is quite simple - if $V_{i n}$ is at ground, the no current will flow through the load $R_{L}$. If $V_{i n} \rightarrow V_{c c}$, the current through the load will rise until

$$
\begin{aligned}
I_{L} & =I_{D S}=\xi\left[V_{c c}\left(V_{c c}-I_{L} R_{L}\right)-\frac{1}{2}\left(V_{c c}-I_{L} R_{L}\right)^{2}\right] \\
& =\frac{\xi}{2}\left[V_{c c}^{2}-I_{L}^{2} R_{L}^{2}\right] \\
I_{L} & =\frac{-1 / \xi \pm \sqrt{1 / \xi^{2}+4 R_{L}^{2} V_{c c}^{2}}}{R_{L}^{2}}
\end{aligned}
$$

If $R_{L} V_{c c} \gg \frac{1}{\xi}$, which implies a large gain for the FET or a high impedance load, this simplifies to

$$
I_{L}=\frac{V_{c c}}{R_{L}}
$$



Figure 5.4: Switching a load using an n-channel MOSFET, with diode production for inductive loads.
which indicates that the FET is working as a perfect switch - when on, the current is simply defined by the load.

The additional components in Fig. 5.4 are included to control various possible failure modes and are not needed in every case. The diode is required for reactive loads (such as motors) which may generate large back-EMF values when turned off. It is not needed for simple resistive loads or loads such as LEDs. The resistor $R_{1}$ is used to limit the inrush current to the capacitor of the MOSFET. Typical values would be between 100 ohms and 1 kohm. Power MOSFETs have relatively large gate capacitance and the inrush current can be large. For smaller MOSFETs, or controls with the ability to source fairly large transient currents, $R_{1}$ is not needed. The resistor $R_{2}$ is used to set a defined switch state if there is nothing connected to the control input and would typically have values of $\approx 1 \mathrm{M} \Omega$. If the input is continuously connected to another device, this resistor is also not necessary.

If a load must be connected to ground, either a negative power supply can be used, which has requires the control signals to also use negative voltages, or a p-channel MOSFET can be used as shown in Fig. 5.5. However, the MOSFETs with the highest current capability are n-channel devices. Also, it is important to note that the logic sense is reversed in the p-channel switching circuit : when the input is at ground, the current flows and when the input is at $V_{c c}$, no current flows.


Figure 5.5: Switching a load using an p-channel MOSFET, with diode production for inductive loads.


Figure 5.6: Increasing the current (power) capacity of a non-inverting amplifier using an npn transistor.

### 5.2.2 Linear Power Boosting

Sometimes a larger current is required from an amplifier circuit than can easily be provided by an op-amp. One solution is use a power op-amp, but the amplification can also be easily provided by adding transistors to the circuit.

If the signal to be amplified is always positive in voltage, the addition of a single power npn transistor as shown in Fig. 5.6 can be sufficient for boosting the power of the op-amp. Through the feedback loop, the op-amp will adjust its output voltage such that the necessary $V_{B E}$ bias voltage is provided, but the current will be supplied primarily by the transistor, providing a current gain of $\beta$ to the circuit. This circuit will not work for bipolar signals, as the npn transistor will only supply current, not sink it.

To handle bipolar signals, two transistors are required in a "push-pull" configuration as shown in Fig. 5.7. Either BJTs or FETs can be used in this circuit. One transistor is responsible for the positive polarity signals while the other is responsible for negative polarity signals. The op-amp handles the bias transition between the two regions (the crossover region) where significant distortions occur in a transistor-only push-pull amplifier.


Figure 5.7: Increasing the current (power) capacity of a non-inverting amplifier using a push-pull stage.

## Chapter 6

## Digital Electronics

The word "digital" is everywhere in modern life : we have digital radio and TV, digital music, digital everything. Many things we never explicitly thought of as "analog" are now proclaimed to be new and digital. What is this power of digital electronics which is sweeping away all existing designs? The power of digital electronics is the accuracy with which it can reproduce exactly the same behavior even under changes in power supplies, external noise, and other factors which can befuddle analog designs with their precise dependence on the values of voltages and currents. Digital designs simplify the state of any given wire into two values: high and low. By dividing the system into just two states, the effects of noise are greatly reduced and this relaxes the design requirements on whole system significantly. With the requirements relaxed in this way, it becomes easy to design very complex circuits and have them work the first time without tuning of component values by hand. These circuits can be integrated onto single chips of silicon and become the microprocessors and digital memory devices which make the digital radio and TV possible.

### 6.1 Combinatorial Logic

The simplification of digital logic into a two-state system means that the mathematical basis changes from a linear (usually) real number structure (voltages) to mathematics based on a very limited number of integers: zero and one. This restricted mathematical system was studied by the mathematician George Boole in the mid-1800s and he defined an algebra of operators and relations between them for studying pure logic problems. These problems had nothing to do with digital computers, but rather were focused on systems of sets where a given item could either be in a given set (one) or not (zero).

As the years went by, Boole's logical work became part of the philosophy of logic rather a part of mathematics, Thus, it was in a 1930s philosophy class that Claude Shannon came across Boole's algebra. Shannon was an electrical engineer who performed his master's work at MIT on a large analog and relay-based "computer" intended for modeling power grids and other challenges of the day. Shannon made the connection between Boole's formalism and the relay systems with which he was working and wrote his master's thesis showing how the application of Boole's formalism could make the design of such a large
system simple. His results were immediately applied to the design of telephone networks and spread widely, carrying Boolean logic to the attention of engineers everywhere.

### 6.1.1 Boolean Logic

Boolean logic is based on a two state system and a set of operators which can combine boolean variables. Boole himself described three operators which have clear analogues in the English language: AND, OR, and NOT. Conventionally, we can describe the behavior of an operator in terms of truth table which indicates the output expected from the system for all inputs. Each operator can also be schematically represented by a symbol and has a textual symbol which is generally used to represent the operator in written text or code. As used in electronics, the operators are frequently referred to as "gates".

## The Basic Gates

The AND operator (or AND gate) matches very closely to the common meaning of "and" in English: the result of the AND operator is true if and only if both (or all) of the inputs are true. The truth table for a two-input AND gate is shown below, along with the schematic symbol for the AND and the textual symbol (\&).

## AND

Schematic symbol:

Textual symbol: \&


Textual symbol:

The second operator introduced by Boole was the OR operator. The English language equivalent of the OR operator is "either A or B or both". This is somewhat different than the concept of "either A or B (but not both)" which is often what is meant in conversation. As we will see below, there is a logic gate which encapsulates this concept: the XOR. The truth table and symbol for the OR gate are shown below.

## OR

Schematic symbol:

Textual symbol:

|

| A | B | $\mathrm{Y}=\mathrm{A} \mid \mathrm{B}$ |
| :---: | :---: | :---: |
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |

The third operator defined by Boole was the NOT operator which is a unary operator which simply inverts the input to get the output. As result, the gate is often called an inverter rather than a NOT gate. The schematic symbol, shown below, is a triangle with a circle or bubble on the output. In digital schematics, this bubble indicates inversion and can be used on the inputs or outputs of a gate depending on the need. Textually, the NOT operator can be written a number of different ways. Common conventions which we will use in this text include an overbar $(\bar{A})$ or a leading exclamations point or tilde $(!A)$.

## NOT

Schematic symbol:

Textual symbol: !


$$
\begin{array}{c||c}
\mathrm{A} & \mathrm{Y}=!\mathrm{A} \\
\hline 0 & 1 \\
1 & 0
\end{array}
$$

## Rules of Algebra

Boolean algebra obeys a set of rules analogous to those of conventional arithmetic.

- Commutative Property The AND and OR are both commutative, which is to say that $A \& B=B \& A$ and $A|B=B| A$.
- Associative Property Both the AND and OR are also both associative, which implies that $(A \& B) \& C=A \&(B \& C)$ and $(A \mid B)|C=A|(B \mid C)$.
- Distributive Property The distributive property is a joint property of two operators which determines how operators can be combined. This property is also available in Boolean algebra, where the AND operator behaves as the multiplication operator is usual arithmetic and the OR behaves as the addition operator.

$$
(A \mid B) \& C=A \& B \mid A \& C
$$

## De Morgan's Theorem

Boolean algebra has another important property, which is named for a major co-worker of Boole's: Augustus de Morgan. De Morgan's theorem states that any logical expression composed of AND, OR, and NOT operators can be equivalently re-expressed by exchanging all AND and OR operators while also inverting all inputs and outputs. The full proof is too long for this text, but we can examine a few cases to convince ourselves of the validity of the theorem.

Let us begin with a simple expression : $Y=A \mid B$. To apply De Morgan's theorem, we must invert all inputs and outputs and exchange OR and AND gates. Thus,

$$
X=![(!A) \&(!B)]
$$

Is $X$ equivalent to $Y$ ? We can write out a truth table to investigate, with a column for each stage in the calculation.

| A | B | $!\mathrm{A}$ | ! B | ! $\mathrm{A} \&!\mathrm{B}$ | !(! $\mathrm{A} \&!\mathrm{B})$ | $\mathrm{A} \mid \mathrm{B}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 0 | 1 | 1 |

On the basis of the truth table, De Morgan's theorem works (at least for OR gates).

### 6.1.2 A Full Set of Gates

While AND, OR, and NOT are a complete set of operators for Boolean algebra, they are not always the most convenient for electronics work. Often, one needs a combination of these gates to perform the necessary logic. For classic digital logic, each type of gate requires its own chip and each chip has a maximum of six gates on it. Thus, a logic calculation using only AND, OR, and NOT can require a large number of chips, even for commonlyused functions. As a result, there are a number of additional gates which are often used to simplify complex logic designs and minimize the chip count.

The most famous of these gates is simply the AND gate with the output inverted, which is called the NAND gate. The schematic symbol for the NAND is the same as for the AND with an inverting bubble on the output of the gate. The NAND is a so-called universal gate as any logic function can be constructed from a collection of NAND gates. Thus, while AND, OR, and NOT are a complete basis for Boolean logic, NAND is both complete and minimal; Figure 6.1 shows how each of NOT, AND, and OR can be constructed from NAND gates. Besides being a universal gate, the NAND is also relatively simple to implement with CMOS transistors, so it has formed as the main building block of integrated digital logic circuits.

## NAND

Schematic symbol:

Textual symbol: $\quad \overline{\&}$ (rarely used)


| A | B | Y = A NAND B |
| :---: | :---: | :---: |
| 0 | 0 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

The NOR gate is the analogous gate to the NAND - an OR gate with its output inverted. Looking at the truth table for the NOR, it is quite clear that the NOR can be considered either as an OR with an inverting output or as an AND with inverting inputs. This duality is simply an application of De Morgan's Theorem. The NOR gate, like the NAND, is a universal gate but was somewhat more difficult to implement than the NAND in early logic families, so the NAND is the most commonly-used universal gate.


Figure 6.1: Implementation of the NOT (a), AND (b), and OR (c) functions using only NAND gates, demonstrating the universal nature of the NAND gate.

## NOR

Schematic symbol:


| A | B | $\mathrm{Y}=\mathrm{A}$ NOR B |
| :---: | :---: | :---: |
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 0 |

There is one more major gate to consider: the exclusive-OR or XOR which implements the logic of "either A or B, but not both" as described above. The XOR can be constructed from combinations of other gates (including a minimum of four NAND gates), but is available as separate device as well. The XOR is one element of an adder circuit capable of adding binary numbers.

Each of the basic gates can be extended in several ways. One extension is the addition of more inputs: the standard "TTL" line of logic chips includes AND gates with up to four inputs (the $74 \times 21$ ) and NAND gates with up to eight inputs (the $74 \times 30$ ). The truth -tables for these devices are simple extensions of the two-input truth tables. The other extension of the basic gates is to cover arrays of bits. Arrays of bits or multi-bit numbers can also be operated on by the AND, OR, and other basic logic operators. The result of such an operation is for each bit of each number to be operated on with its partner in the other number. For an AND operation between two numbers $A$ and $B$, the lowest bit of $A$ is ADDed with the lowest bit of $B$, the second bit of $A$ with the second bit of $B$ and so on. This allows us to extend boolean logic to arrays of bits or boolean numbers.

## Special Topic: Binary Numbers

Basic boolean algebra allows only two numbers: zero and one. However, we can express larger numbers by using binary notation. To understand binary notation, it is important to recall the meaning behind decimal notation : each digit represents an additional power of ten. Thus, we can decompose any number into a sum of powers of ten each multiplied by a number between zero and nine.

$$
4804=4 \times 10^{3}+8 \times 10^{2}+0 \times 10^{1}+4 \times 10^{0}
$$

With binary numbers, we are allowed multipliers of zero or one, so we must represent digits as sums of powers of two:

$$
\begin{aligned}
4084 & =4096+512+128+64+4 \\
& =1 \times 2^{12}+1 \times 2^{9}+1 \times 2^{7}+1 \times 2^{6}+1 \times 2^{2}
\end{aligned}
$$

Every position in the binary number which is not one must be zero, so we see that 4084 is 1001011000100 in binary. Binary numbers require a lot more digits than decimal to represent the same value, but binary numbers have the major advantage of being amenable to manipulation by digital logic and thus are the basis of all the computations done inside a modern computer or calculator.

Conversion from binary to decimal is straightforward: simply multiply each digit by the appropriate power of two and add up the result. Conversion from decimal to binary can be more challenging, but one method is the division and remainder trick. Repeatedly divide the decimal number you wish to convert to binary by two and record the remainder each time. Stop when you are left with zero. Then the binary number is the list of remainders starting with the first division's remainder at the least-significant bit.

For example, let's convert 53 to binary this way.

$$
\begin{aligned}
& \text { 53/2 }=26 \text { rem }=1 \rightarrow \text { right-most digit } \\
& 26 / 2=13 \text { rem }=0 \\
& 13 / 2=6 \text { rem }=1 \\
& 6 / 2=3 \text { rem }=0 \\
& 3 / 2=1 \text { rem }=1 \\
& 1 / 2=0 \quad \text { rem }=1 \rightarrow \text { left-most digit }
\end{aligned}
$$

Thus, $53=110101=2^{5}+2^{4}+2^{2}+2^{0}=32+16+4+1$. Check!

## XOR

Schematic symbol:


| A | B | $\mathrm{Y}=\mathrm{A}^{\wedge} \mathrm{B}$ |
| :---: | :---: | :---: |
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

Textual symbol:

## Example

What is the result of 4 AND 5 if both 4 and 5 are three bit numbers? First, we need to convert the two numbers to binary.

$$
\begin{aligned}
& 4=1 \times 2^{2}+0 \times 2^{1}+0 \times 2^{0}=100 \\
& 5=1 \times 2^{2}+0 \times 2^{1}+1 \times 2^{0}=101
\end{aligned}
$$

Now, we can apply the AND operator to each bit individually: | 4 | 1 | 0 | 0 |
| :---: | :---: | :---: | :---: |
|  |  | $\&$ | $\&$ |
| 5 | $\&$ |  |  |
| 5 | 1 | 0 | 1 |
|  |  | 1 | 0 | 0

So the result of 4 AND 5 is $100=4$.

### 6.1.3 Decoders and Multiplexers

The basic logic gates are mathematically sufficient for any purpose, but a large number of logic gates may be needed for some common operations. These common operations, with enough use, have been become codified as standard operations in their own right. Two of the most important such higher-level gates are the decode and the multiplexer. Both functions involve operations using arrays of bits or binary numbers and are generally not directly represented by simple symbols in text, given the large number of inputs and outputs from the devices.

The term "decoder" is used for a number of different devices, but in this case we are interested in the version of the device with $n$ inputs and $2^{n}$ outputs. The decoder accepts a binary number as input and produces a HI value on the output with the index number of the input binary number. All other outputs will be low.

Considering the simple 2 -to- 4 decoder, it is relatively simple to write down the equations for all four outputs in terms of the inputs:

$$
\begin{aligned}
& Y_{3}=S_{1} \& S_{0} \\
& Y_{2}=S_{1} \&!S_{0} \\
& Y_{1}=!S_{1} \& S_{0} \\
& Y_{0}=!S_{1} \&!S_{0}
\end{aligned}
$$

## 2-to-4 Decoder

Schematic symbol:


| $S_{1}$ | $S_{0}$ | $Y_{3}$ | $Y_{2}$ | $Y_{1}$ | $Y_{0}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 |

A simple-minded implementation of the 2-to-4 decoder would require four AND gates and two NOT gates, while a three-input decoder would require sixteen AND gates if we limited ourselves to the two-input variety! Given the utility of the decoder, it is not surprising that the whole circuit can be found integrated into a single chip.

The inverse function of the decoder is the encoder, which produces a binary number to indicate which of several inputs is active. When more than one input is active, the behavior of different types of encoders varies. The most common encoder will produce an output equal to the highest input which is active. Such a device is called a priority encoder.

The multiplexer is a circuit which allows one to make a choice between signals. The multiplexer has two sets of inputs : the data inputs and the control bits which select between the multiplexer inputs. The multiplexer sets its output to the value of the data input chosen by the control bits, as shown in the truth table below. This truth table introduces a new character to the truth table : $X$. The symbol $X$ means "don't-care"; the output of the logic circuit will not depend on whether the input is high or low. This convention allows the multiplexer truth table to be expressed with a mere eight rows rather than the 64 rows which would otherwise be required.

## Four-input multiplexer

Schematic symbol:


| $S_{1}$ | $S_{0}$ | $D_{3}$ | $D_{2}$ | $D_{1}$ | $D_{0}$ | $Y$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | $X$ | $X$ | $X$ | 0 | 0 |
| 0 | 0 | $X$ | $X$ | $X$ | 1 | 1 |
| 0 | 1 | $X$ | $X$ | 0 | $X$ | 0 |
| 0 | 1 | $X$ | $X$ | 1 | $X$ | 1 |
| 1 | 0 | $X$ | 0 | $X$ | $X$ | 0 |
| 1 | 0 | $X$ | 1 | $X$ | $X$ | 1 |
| 1 | 1 | 0 | $X$ | $X$ | $X$ | 0 |
| 1 | 1 | 1 | $X$ | $X$ | $X$ | 1 |

The four-input multiplexer can be realized in logic by combining a 2 -to- 4 decoder with four AND gates and a four-input OR gate. The decoder will provide the selection signal
which is combined with the input signals to produce the output. Real integrated multiplexers are often built with a simpler design which replaces the AND-OR tree with pass-gates which act as simple switches. The inverse function of a multiplexer is a demultiplexer, which is quite similar in design to a decoder.

### 6.1.4 Memory: The Universal Gate

The multiplexer, decoder, and similar devices provide a higher level of abstraction between the low-level gates and the logic of the design. However, one often must still add the odd inverter or other discrete gate to a design even with these high-level constructs available. This arises because the individual logic elements are insufficiently flexible to cover every situation. The solution is to use a programmable device, which will be discussed in more detail in Chapter 7. One of the main techniques in programmable logic is to replace the fixed logic elements with read-only-memories or ROMs.

A memory device is a simple extension of the ideas used by the multiplexer. The control lines have the same function: to select the input which will be connected to the output, but the inputs are not external wires but rather internally-saved values. These data values are programmed initially by the circuit designer (in one of several ways which will be discussed later), but afterward they are fixed. The memory thus behaves like an array of bits, with the output bit chosen from the array on the basis of the control inputs. For a memory, the selector bits are usually referred to as address bits, as they provide the address within the array of bits.

Such a memory can be used to define any possible combinatorial logic with a number of inputs equal to its address inputs. The designer need only determine what the output is for each possible input combination and program the appropriate value into each memory location. Most memories allow the storage of more than one bit at each location, which allows for multiple parallel logic functions of the same inputs to be performed by the memory. This flexibility allows a ROM to behave as a generic logic function generator.

## Example

Convert the complex three-input, two-output circuit below into the data to be loaded into a eight-by-two ROM (with address inputs $A_{2}, A_{1}, A_{0}$ and data outputs $D_{1}$ and $D_{0}$.


First, let us work out the complete truth-table for this complex circuit.

| $B$ | $C$ | $E$ | $W$ | $Y$ |
| :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 0 |

Next, we must assign the inputs of the circuit to address lines and the outputs to data bits. This choice is arbitrary and up to the designer - one must simply be consistent! For this example, we choose $B=A_{2}, C=A_{1}, E=A_{0}$ for the inputs and $W=D_{1}, Y=D_{0}$. Now, we can convert the truth table into an address map of values to write into the memory. For convenience, we can write the values in both binary and decimal.

| Binary <br> Address |  | Decimal |  |
| :---: | :---: | :---: | :---: |
| Address | Data |  |  |
| 000 | 11 | 0 | 3 |
| 001 | 01 | 1 | 1 |
| 010 | 10 | 2 | 2 |
| 011 | 11 | 3 | 3 |
| 100 | 11 | 4 | 3 |
| 101 | 11 | 5 | 3 |
| 110 | 11 | 6 | 3 |
| 111 | 10 | 7 | 2 |

### 6.2 Logic Standards

Up to this point, we have dealt with digital logic essentially as pure math: no implementation of the circuitry has been discussed. In fact, the details of implementation for digital logic are quite often left unknown to the circuit designer: the manufacturer of a chip may change the internal implementation at will. A circuit designer need not know the details of the implementation, but two parts of the specification of the logic element are critical: the truth-table implemented by the device and the logic levels which are accepted and driven by the device.

The logic level parameters define what each device will accept as an input for a logic low and a logic high and what it agrees to drive on its output as logic low and high. This is the point where the realities of analog electronics intrude somewhat into the perfect mathematical world of digital logic. In reality, all these digital bits must be represented as voltages, currents, or charges and the gates must be designed with sufficient noise immunity or else a low will be confused for a high or vice versa.

### 6.2.1 Positive Voltage Logic Families

In the early years of digital logic, several semiconductor device companies pushed different standards for logic levels and device conventions. Texas Instruments produced a major winner in this battle with its introduction of the 7400 -series of chips based on transistor-transistor-logic or TTL. The 7400 series of chips included a full set of logic functions, both combinatorial and sequential, and included a number of helpful conventions including the standard placement of the power and ground pins on the package. The 74 -series of chips has been a runaway success, and many other manufacturers have produced chips following the pinout and specifications of the TI chips. Additionally, as new logic standards are developed new 74 -series chips are generally produced following the old pinout but the new logic standards. These differing logic standards or performance characteristics are indicated by letters included in the middle of the chip specifier. For example, the 7402, 74F02, 74LS02, and 74AHCT02 are all quad two-input NOR-gate devices, but represent the original (obsolete) TTL, "fast" TTL, "low-power Schottky" TTL, and "advanced highspeed CMOS with TTL-level inputs", respectively.

Despite the alphabet soup involved in these different 74 -series chips, they all share a few common characteristics. The logic levels in these chips are defined by voltages rather than currents. A logic low is indicated by a voltage near ground, while a logic high is indicated by a positive voltage above ground, usually near the positive power supply which was 5 V in classic TTL, but is switching to 3.3 V or lower in modern devices. It is important to notice that each logic family has two standards for low logic levels and two for high. One is a constraint on the output value associated with a given logic level and the other is a constraint on the input level which is accepted for a given logic level. These two levels are separate because the impedance of a TTL input is not infinite (particularly at high frequency) and the wires which carry the signal from the logic driver to the logic input have their own impedance and noise. Thus, each logic family has a highest allowable output voltage for a logic low $\left(V_{O L}\right)$ and at a somewhat higher voltage is the highest input


Figure 6.2: Detail of the protection diodes included on modern digital logic inputs.
voltage which should be considered as logic low $\left(V_{I L}\right)$. On the logic high side, the minimum allowable output voltage for a logic high $\left(V_{O H}\right)$ is higher than the minimum input voltage considered as logic high $\left(V_{I H}\right)$. The differences $V_{I L}-V_{O L}$ and $V_{O H}-V_{I H}$ represent noise and resistance immunity - a noise pulse of less than $V_{I L}-V_{O L}$ will not alter in any way the interpretation of the value as low.

|  | 5 V |  |  |  | 3.3 V | 2.5 V |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 74 | 74 LS | 74 HC | 74 HCT | 74 V | 74 ALVC |
| $V_{O H}$ | 2.4 V | 2.7 V | 4.8 V | 4.8 V | 2.6 V | 2.2 V |
| $V_{I H}$ | 2 V | 2 V | 3.2 V | 2 V | 2 V | 1.7 V |
| $V_{I L}$ | 0.8 V | 0.8 V | 1.4 V | 0.8 V | 0.8 V | 0.7 V |
| $V_{O L}$ | 0.4 V | 0.5 V | 0.2 V | 0.2 V | 0.5 V | 0.4 V |
| Gate propagation | 12 ns | 18 ns | 10 ns | 11 ns | 8 ns | 3 ns |

Table 6.1: Table of logic family thresholds for a selection of different positive-voltage logic standards.

When connecting together devices of different families, care must be taken. Table 6.1 lists a number of different logic families as well as the input and output thresholds for each family. For logic family $A$ to be able to drive $\operatorname{logic}$ family $B$,

$$
\begin{aligned}
V_{O H}^{A} & \geq V_{I H}^{B} \\
V_{I L}^{A} & \leq V_{I L}^{B}
\end{aligned}
$$

Thus, 74 HC devices are able to drive signals to 74 LS devices ( $4.8 \mathrm{~V} \geq 2 \mathrm{~V}$ and $0.2 \mathrm{~V} \leq 0.8 \mathrm{~V}$ ), but 74LS devices cannot drive 74 HC devices since $2.7 \mathrm{~V}<3.2 \mathrm{~V}$. In such a situation, one must use an interfacing technique. For example, when connecting 74LS logic to 74HC logic one may use a layer of 74 HCT chips, which as seen in Table 6.1, have LS-compatible input levels and HC-compatible output levels.

Besides the minimum output high voltage levels, today one must be careful with the maximum output high voltage logic levels. From the introduction of TTL until the mid1990s, 5V logic reigned supreme, but with higher logic densities chip designers have switched to lower power supply voltages to limit the power dissipation of their designs. This trend began with a switch to 3.3 V logic, but 2.5 V or even lower voltage logic is common today. When newer logic devices are connected to older 5 V systems, care must be used. Most modern logic chips include protection diodes as seen in Figure 6.2 which are intended to protect the sensitive transistors of the chip from electrostatic discharge and other damage. When these chips are used with others of the same power supply level, there is no problem - the logic levels are never high enough (or low enough) to allow either of the protection diodes to go into conduction. However, a 5V logic signal can easily turn on the protection diode and send a potentially very large current into the input, through the diode, and back out to the (lower voltage) positive supply. This situation can cause damage to the device. The solution depends on the speed and numbers of signals in question; a few slower signals can easily be "converted" using a resistive voltage dividers. For some chips, it is sufficient to include a series resistor to limit the current flowing through the protection diode. Large numbers of signals may make use of a level-translator chip the more convenient option.

### 6.2.2 Shared Lines



Figure 6.3: Schematic representation of a push-pull or totem-pole digital logic output.

In a complex digital system, there are often many logic devices which must be connected together. Often, there are connections which must be shared: either connections between the devices or connections to an outside resource. In theory, one could provide direct, dedicated logic connections for all purposes but this solution is often unworkable or expensive. Instead, when the parts of a system do not need to continuously exchange
information they can share a set of wires for multiple purposes. These sets of shared wired are called buses and we will discuss the concepts of buss-es in more detail in later chapters. Buss-es tend to share a couple of technical needs: lines which can be used to transfer binary data to and from nodes on the bus and also control lines which can safely be used by any-of-many devices.

To understand the issues involved in shared lines, it is useful to look in some detail at the output of a discrete logic gate. At the output of the gate, the result must be driven onward to the next chip. This can be done by a "totem-pole" driver such as that in Figure 6.3. The driver consists of two transistors and an inverter. If the $A$ value is high, the upper transistor turns on and connects the output $Y$ to the positive power supply while the lower transistor will be off. If the $A$ value is low, the upper transistor will turn off and the lower transistor will turn on - connecting the output to ground.


Figure 6.4: Schematic representation of a tristate logic buffer where the $T$ input can be used to push the output to high-impedance.

The binary data is exchanged in a bussed system using a set of data lines - any of several devices is allowed to set values on the bus. However, if two devices try to assert different values on the bus at the same time, the result is bus contention - the output stages of the two devices "fight" to determine what state the contended lines will end up in: the upper transistor of one driver will be sourcing large amounts of current into the lower transistor of the other. To avoid this situation, buses generally use tristate logic drivers. These are logic drivers which can be in one of three states: driving 0 , driving 1 , or not driving (highimpedance). Figure 6.4 shows how this is achieved. When the added $T$ input is low, the circuit behaves as before - a high input turns on the upper transistor and a low input turns on the lower transistor. If the $T$ input is high, however, both transistors turn off and the output is not connected to either ground or the positive supply by a low-impedance path. Thus, the tristate input can be used to disable the output of the gate. By using a decoder or other control signals to enable just one device to output on the bus at a time, the data lines


Figure 6.5: Schematic representation of an open collector output driver. The pull-up resistor is usually provided externally to the device.
can be shared between many devices.
For control purposes, it can be very useful to have a single line which can be safely driven by multiple chips without having to provide enable logic. These lines are often used to indicate "error" or "action needed". However, without the explicit enable logic, the tristate technique cannot be used. Instead, the common solution is a device where the upper transistor of the totem-pole is eliminated as in Figure 6.5. In this case, the driver can pull the output low but cannot pull it high. Instead, a moderately large-valued ( $\sim 3 \mathrm{k} \Omega$ ) external resistor is connected to the line. This resistor serves to weakly pulls up the line. When no driver on the line is pulling it low, the resistor will cause it to drive high, but the output impedance of the activated digital driver is much less than $1 \mathrm{k} \Omega$ so the line will be low. This type of logic is often called "open-collector" or "open-drain" logic, indicating that the output driver does not include the active high drive.

### 6.2.3 Differential and Specialized Standards

The positive voltage-level standards are not the only logic standards in use. For noisy environments, the high impedance of the TTL-style inputs is not necessarily ideal - small induced currents on the lines can easily produce large voltage spikes on the inputs. As signaling speeds increase, TTL tends to produce its own noise - the large voltage switch between low and high induces noise on neighboring lines. Standard TTL is not appropriate for long ( $>10 \mathrm{~m}$ ) cable runs unless large efforts are made to control noise and is generally limited to signaling speeds below 100 MHz . For higher speeds and longer cable runs, a number of different signaling standards exist, a sample of which are listed in Table 6.2.

Most of the standards are quite similar and involve the use of two lines rather than one. A logic level is determined by the relative voltage levels of the two, or equivalently by the direction of current flow around a loop along one line and back on the other. The differential technique has much the same effect as in analog logic: common-mode noise is

|  | NIM | ECL | PECL | LVPECL | LVDS | HSTL |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Power supply | -2 V | -5.2 V | 5.0 V | 3.3 V | 3.3 V | 1.5 V |
| Logic high | -20 mA | -0.9 V | 4.0 V | 2.5 V | 1.4 V | 1.15 V |
| Logic low | 0 mA | -1.75 V | 3.2 V | 1.5 V | 1.0 V | 0.35 V |
| Differential? | No | Usually | Yes | Yes | Yes | Often |
| Termination | $50 \Omega$ | $100 \Omega$ | $100 \Omega$ | $100 \Omega$ | $100 \Omega$ | $100 \Omega$ |

Table 6.2: Table of logic family thresholds for a selection of current-based, negative voltage, and differential logic standards.
canceled leaving only the desired signal. Additionally, the differential standards have low characteristic impedances which limits the impact of current-induced noise.

### 6.3 Sequential Logic

Combinatorial logic allows us to define logical conditions based on input variables, but these values are always the "current" ones. Often, it is convenient to have some history or memory in the system as well - that way the current behavior depends on both previous settings and new inputs. It is also very useful to segment calculations in time - this aids with inevitable problems of combinatorial delay and glitch conditions. The solution is to add a device capable of retaining the history as well as a pace-setter for the calculations: a system clock.

### 6.3.1 Limits of Combinatorial Logic

Combinatorial logic is not infinitely fast - each gate requires some time to act on changes to its input. A gate's propagation time depends on the logic family (see Table 6.1 for some propagation time values), but also depends on temperature and electrical load. When many gates are connected together, this can lead to temporary unexpected conditions.

Consider the circuit in Figure 6.6. This is a simple AND+OR combination where one signal $(C)$ goes directly to the OR gate and two ( $A$ and $B$ ) are AND-ed together to create an intermediate term $E$. Despite the simplicity of the circuit, it can generate unexpected results depending on the ordering of changes to the inputs. A useful tool for visualizing the effects of ordering is the timing diagram. Each signal is drawn as a trace, similar to that on a oscilloscope. Signals are separated from each other vertically, while time proceeds along the horizontal axis. Let us use this tool to visualize the effect of the following sequence of events:

1. Initially, $A=0, B=1$, and $C=1$.
2. At $t=0, A$ goes from 0 to 1 (a rising edge).
3. At $t=2 \mathrm{~ns}, C$ goes from 1 to 0 .


Figure 6.6: Example circuit for glitch-analysis.

First, consider this if the elements have no propagation delays.

1. $A \& B=E=0$, but $E|C=0| 1=1$ so the output is high.
2. When $A$ goes from 0 to 1 , then $E \rightarrow 1$ immediately. $E|C=1| 1=1$ so the output remains high.
3. When $C$ goes from 1 to $0, E|C=1| 0=1$ so the output remains high.

Thus, the output does not change from being high in the case of infinitely fast logic.
However, let us consider the result if each gate has a propagation delay of 10ns. In this case, the result will appear on the output of any gate 10 ns after a change on the input. The timing diagram for this situation is shown in Figure 6.7. The addition of the propagation delay makes all the difference in the results.

1. Initially, the output is high as the inputs have been stable for a long time.
2. When $A$ goes from 0 to 1 , then $E \rightarrow 1$ after 10 ns. However, as $C$ is still high, $Y=1$.
3. When $C$ goes from 1 to 0 , nothing happens immediately - the output remains the same.
4. At $t=10 \mathrm{~ns}, E \rightarrow 1$. However, the OR gate does not act immediately either.
5. At $t=12 \mathrm{~ns}, Y \rightarrow 0$ because 10 ns ago $C=0$ while $E=0$.
6. At $t=20 \mathrm{~ns}, Y \rightarrow 1$ as the OR gate reacts to the rising edge on $E$.

Compared to the infinitely-fast case, the realistic case has an unexpected change in the output value which appears momentarily before disappearing again. Such a transient change is


Figure 6.7: Timing diagram analyzing the behavior of the circuit in Figure 6.6.
called a glitch and glitches can wreak havoc in digital logic systems - the short, unexpected pulses can radiate through an interconnected web of logic spawning additional glitches.

Glitches can be reduced by careful design of circuits to avoid situations where signals have different propagation delays, but such tuning is of limited effectiveness. The propagation delays change as a function of temperature, supply voltage, and output load so the precision of any tune is limited. Instead, designers have switched to the use of sequential logic where glitches are tolerated - as long as they work themselves out before the next rising clock edge.

### 6.3.2 D Flip-Flop

The heart of sequential logic is the flip-flop. There are a number of types of flip-flop, but we will focus on the most widely-used flip flow: the D Flip-Flop. The simple D flip-flop acts as single bit of memory and has two inputs. One is the data or $D$ input and the other is the clock input. The behavior of a D flip-flop is that the output will be constant except when the clock has a rising edge. At that point, the value on the $D$ input is copied to the $Q$ output and remains constant there until the next rising edge of the clock.

D flip-flops can also be obtained with additional SET and/or RESET inputs. These resets force the flip-flop to either true (SET) or false (RESET) without waiting for a clock. Generally, these inputs are active-low and used for special conditions where immediate action is needed.

## D Flip-Flop

Schematic symbol:


| $D$ | Clock | $Q$ | $\bar{Q}$ |
| ---: | :---: | :---: | :---: |
| X | 0 | $D^{\prime}$ | $!D^{\prime}$ |
| X | 1 | $D^{\prime}$ | $!D^{\prime}$ |
| 1 | $\uparrow$ | 1 | 0 |
| 0 | $\uparrow$ | 0 | 1 |

### 6.3.3 Counters

The D flip-flop (or register) is a key building block in sequential logic. One of the simplest tasks in sequential logic is counting - keeping a record of the number of rising edges observed on a line. Counters (devices which count) are critical items in many measurements and form the core of unexpected devices such as analog-to-digital and digital-to-analog converters.

To understand the circuits needed to implement a counter, it is useful to write down what the counter is expected to in the proper form: binary numbers. Let us write down the counter sequence for a three-bit counter - one which should be able to count to seven.

| Decimal | Binary |
| :--- | ---: |
| 0 | 000 |
| 1 | 001 |
| 2 | 010 |
| 3 | 011 |
| 4 | 100 |
| 5 | 101 |
| 6 | 110 |
| 7 | 111 |

When the counter reaches seven, it will return to zero again, a condition called wrapping.
Notice that the lowest bit is constantly toggling - low/high/low/high. The next bit toggles whenever the lowest bit transitions from high to low. The highest bit similarly transitions when the bit below it goes from high to low. This suggests a design where each bit is represented by a D flip-flop with $\bar{Q}$ connected back to $D$. The lowest bit would use the signal to be counted as its clock and each level higher would use the $\bar{Q}$ output of the bit below it. Such a design is shown in Figure 6.8 and is called a ripple counter.

The ripple counter is so-called because increments ripple through the bits - each bit affecting the next. This can be clearly seen in a timing analysis of the ripple counter as seen in Figure 6.9. The lowest bit clearly toggles high/low with a small propagation delay from the input signal. The next highest bit toggles when the lowest bit has a falling edge, but with an additional propagation delay since the output of the lower bit is connected to the clock input and the process of transferring the input value to the outputs only begins when


Figure 6.8: Design of a 3-bit ripple counter.
the clock ( $\overline{C[0]}$ in this case) has a rising edge. As a result, there is a moment when one is followed by zero before resolving into two.

This effect may or may not be a problem, depending on the use to which the value is being put. If it is vital that the counter always count sequentially (with no glitches), this design is unacceptable. However, many designs may not care about transient counter values - the more important question is the number at the end. The ripple counter does give a valid count in the end, though it may require some waiting for a large ripple counter to become valid. The extreme simplicity of the ripple counter has much to recommend it and the ripple counter is effective up to very high count rates.

The alternative counter design is a completely synchronous one where the clock for all bits is the same: the input signal $A$. The design for the synchronous counter requires significantly more logic. We can begin by noticing that the behavior of the lowest bit is already ideal - its design will not change. We expect each bit of the counter to be a function of its own value and the values of those below it. Next, we can write out the truth table for $Q_{1}(n+1)$ as a function of $Q_{0}(n)$ and $Q_{1}(n)$ - this is simply the counting sequence.

| $Q_{1}(n)$ | $Q_{0}(n)$ | $Q_{1}(n+1)$ | $Q_{0}(n+1)$ |
| :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 |



Figure 6.9: Timing analysis of a 3-bit ripple counter.

The truth table clearly indicates that $Q_{1}(n+1)=Q_{1}(n) \operatorname{XOR}_{0}(n)$.
With this success, we can continue to the highest bit $Q_{2}$ which is rather more complicated.

| $Q_{2}(n)$ | $Q_{1}(n)$ | $Q_{0}(n)$ | $Q_{2}(n+1)$ |
| :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 |

To determine the necessary logic, we can split into two cases: when $Q_{2}(n)$ is low and when it is high. When $Q_{2}(n)$ is low, the result is $Q_{1}(n) \& Q_{0}(n)$ while when $Q_{2}(n)$ is high, we have $Q_{1}(n)$ NAND $Q_{0}(n)$. There are several ways to implement this result, but we will choose the direct implementation of our result:

$$
Q_{2}(n+1)=!\left(Q_{2}(n)\right) \& Q_{1}(n) \& Q_{0}(n) \mid Q_{2}(n) \&!\left(Q_{1}(n) \& Q_{0}(n)\right)
$$

The resulting circuit is shown in Figure 6.10. The logic for the higher bits becomes significantly more complicated (though it could be simpler than shown here - see Problem 2). This means that the combinatorial logic grows for each successive bit and sets an upper limit on the speed of a synchronous counter depending on the number of bits in the counter and the propagation delays in the "carry logic".

### 6.4 State Machines

A state machine is way of organizing sequential behavior in a system or a small part of a system. In a state machine, the behavior is divided into a series of cases or states and


Figure 6.10: Implementation of 3-bit synchronous counter.
rules are given to control the transition of the state machine from one state to another. The logical construct of the state machine is useful for organizing many tasks in digital electronics and system design. The state machine formalism helps structure the design and indicates problems which may exist.

The design of a state machine begins by enumerating the states in question and the conditions which cause a transition. Consider as an example the following system - a simple scalar:

- The system is required to count the number of rising edges (pulses) during the time that a gate input is high.
- When the gate input goes low, the system should present the number of observed pulses on its output and raise a flag indicating completion.
- The external data acquisition computer (DAQ) will read the output and then send a pulse to reset the scalar. The output and flag should then be zeroed and the scalar should wait for a new gate.

The state machine in this design is not the counter itself : that is an element controlled by the state machine. The state machine is controlling element, but is generally simple. It usually consists of a small number of registers, with somewhat complex logic feeding the register changes.

The conversion of this abstract state machine into a implemented state machine begins by determining the states of the system based on the desired behavior.

1. IDLE - The system is waiting for a gate. When the gate arrives, the system transitions to COUNTING.
2. COUNTING - The system is counting pulses. When the gate stops, the system transitions to OUTPUT.
3. OUTPUT - The system is displaying the output values. When the reset pulse arrives, the system transitions to IDLE.

These transitions will occur depending on two input signals.

1. GATE - As long as the GATE signal is high, the system should count.
2. RESET - Once the DAQ has read the value, it will send the reset value.

We can now write down a decision or control tree which specifies how the system goes from state to state.

| Initial State | Condition | Next State |
| :--- | :--- | :--- |
| IDLE | GATE high | COUNTING |
|  | GATE low | IDLE |
| COUNTING | GATE high | COUNTING |
|  | GATE low | OUTPUT |
| OUTPUT | RESET high | IDLE |
|  | RESET low | OUTPUT |

We can now encode the state machine into hardware by selecting a representation of the state as registers. The state can be encoded as binary numbers in the usual counting sequence, as binary numbers in a "Gray code" where only one bit changes between each pair of state, or in a number of different ways. For this example, we select a simple encoding called "one-hot" encoding which appropriate when the number of states is small. In this encoding, each state is represented by one register and the design must guarantee that one and only one register is on at any time.

In the one-hot encoding, we must write down the conditions for the next state (after the next clock edge) to be each of three states.

$$
\begin{aligned}
\mathrm{IDLE}^{n+1} & =\mathrm{IDLE}^{n} \&!\mathrm{GATE}^{2} \mathrm{OUTPUT}^{n} \& \mathrm{RESET} \\
\mathrm{COUNTING}^{n+1} & =\mathrm{IDLE}^{n} \& \mathrm{GATE}^{\mathrm{COUNTING}} \& \mathrm{CATE}^{n} \\
\mathrm{OUTPUT}^{n+1} & =\mathrm{COUNTING}^{n} \&!\mathrm{GATE} \mid \mathrm{OUTPUT}^{n} \&!\mathrm{RESET}
\end{aligned}
$$

These conditions describe simple AND-OR terms for each of the state machine bits and can be implemented using usual logic. We must arrange using set/reset logic that the IDLE register is selected at power-on and none of the others is on. This is a common initialization problem which often appears at power on and particular care must be taken to achieve the necessary initial conditions.

Once the state machine structure is determined, we can then determine how to generate any of the necessary output signals. For this design, we expect three output signals to control the counter and interface with the external DAQ.

1. COUNT - Enables the counter to increment
2. FLAG - Data ready flag
3. CLEAR - Resets the counter

Each of these outputs can be directly associated with a single state. During IDLE, the CLEAR can be always asserted. During COUNTING, the COUNT output should be high. During OUTPUT, the FLAG must be set. Thus, with the one-hot encoding we simply connect the appropriate register outputs to the counter controls and the external DAQ. If we had used binary encoding, we would have to have used combinations of the state bits to drive the outputs.

## Problems

1. Determine the truth table for the circuit shown below. Find an implementation for the final output using only NAND gates (it can be done in with just three gates).

2. Express the $Q_{2}(n+1)$ condition for a three-bit synchronous counter using a single XOR gate and a single AND gate. What does your result suggest for extension to higher bits?
3. Complete the timing diagram for the circuit below for the given inputs ( $A$ and $B$ ). The clock frequency is $10 \mathrm{MHz}(\mathrm{T}=100 \mathrm{~ns}$ ) and the propagation delay through any gate or flip-flop can be assumed to require 10 ns . Determine the difference in behavior (if any) if the length of the pulse on $A$ is shortened to 20 ns .


## Chapter 7

## Programmable Digital Logic

Computers and portable music players clearly demonstrate the power of digital electronics, but they contain relatively few standard gates. A computer motherboard may contain a handful of individual gate chips as "glue logic", but most of the work in the device is done by general purpose logic whose behavior is controlled by a program of one sort or another. The availability of these general-purpose programmable devices, combined with computer design support, explains the high rate at which new digital devices appear on the market. The circuits themselves are very simple - all the work is in the programming and software tools exist to make that relatively simple.

While exceptions and mixtures exist, we can divide programmable digital electronics into two major families. The first group consists of devices which attempt to provide standard digital logic functions in a generic form. On an ideal level, these devices appear as an array of generic gates which can be tied together and specialized for any particular purpose by the program which loaded into it. The other group of devices are sequential data processing devices whose programs are series of instructions which are used one by one to select the next instruction to execute and the changes to made in the memory of the device. This group contains the microprocessors which form the heart of a personal computer. It also contains much simpler and cheaper devices called microcontrollers which can be found in many household items.

### 7.1 Programmable Logic Devices

Programmable logic devices are the direct logical extension of the general techniques of digital logic. In general, PLDs consist of a network of logic and registers whose interconnections are controlled by the program downloaded into them. The technology in use has changed and developed over the years.

### 7.1.1 PALs and CPLDs

The first major family of programmable logic devices was the programmable-array-logic (PAL) devices originally introduced by Monolithic Memories. These devices were quite
successful and were widely copied by other manufacturers, though often using different acronyms such as Programmable Logic Array (PLA) or Generic Array Logic (GAL).

The design for a simple two-input/one-output PAL is shown in Figure 7.1. The PAL design is based on AND-OR terms. Each output value is determined by the OR of several intermediate terms, each of which is the output of an AND. The inputs to the AND are chosen from the inputs to PAL and the inversions of those inputs, as seen in the figure. As shipped from the factory, all combinations in the PAL are connected, which provides a relatively useless device. The connections between the inputs and the AND gates are made by way of a circuit element which can be modified from the outside - originally these were small fuses. The user of the PAL must decide which fuses to blow in order to define the necessary logic. The PAL contains internal pull-up resistors which mean that any signals which have their fuses blown will be replaced by HIGH values, and thus will not affect the AND logic.

The device pictured in Figure 7.1 is a relatively sophisticated PAL, as it includes a programmable choice for the output between a pure combinatorial output and a registered output. This choice is made by way of a multiplexer at the output stage. As shown, the default behavior is combinatorial but by blowing the fuse to ground, the output can be taken from the flip-flop. Notice that when the output is used in combinatorial mode, the flipflop is present (and even active, if the clock is present), but it is ignored. The presence such unused registers and chip capacity is a common effect in programmable logic devices - the devices are designed for "all" possible cases, and any specific case will leave some resources unused.

The term "blow the fuse" is an accurate description of the mode of programming for early PALs. The devices were placed into special programmers which accepted a list of fuses to blow from a computer and then special voltage combinations were applied to the chip's pins to select and then blow the chosen fuses. Once programmed, the PAL could not be erased and reused - additional fuses could be blown, but none replaced. These devices were referred to as One-Time-Programmable (OTP). The time required for programming the devices could be significant for large productions quantities, so manufacturers also provided options to create a custom chip mask to apply the fuse selection during the production of the chip itself - after using the conventional OTP devices for the development process. Later devices (particularly the GAL devices), replaced the fuses with non-volatile memory bits. Once burned in, the program was stable but the device could be returned to its initial state either by the application of special erasure voltages or by exposing the silicon to ultraviolet light through a quartz window on the package.

Classic PAL devices allow all inputs to be combined for any (or all) outputs - a design which is highly flexible but which cannot be easily scaled up as the pin count of the device increases. It is generally rare that all the inputs are needed for any output and the waste of resources for a 100-input/100-output PAL (compared to four $25 / 25$ PALs) would be very large for most designs. As a result, the basic PAL design is not scaled to large pin counts. Instead, larger devices use a architecture where each block of roughly sixteen inputs is connected to a PAL-like block. Some outputs of the block connect to output pins, but there are additional outputs which lead to an interconnection matrix. Each block also accepts some inputs $(\mathrm{O}(8))$ from the interconnection matrix. The matrix allows some signals to be shared


Figure 7.1: Programmable AND-OR tree as used by Programmable Array Logic/Generic Array Logic devices.


Figure 7.2: Structure of a Xilinx Coolrunner-II CPLD. The "PLA" is the PAL-like logic block while the "AIM" is the interconnection matrix between the logic blocks. Distribution of clocks and global resets is also shown.
between the blocks of inputs and represents a trade-off between the fully-connected PAL design and many individual devices. Devices of this type are called Complex Programmable Logic Devices (CPLDs) and an example is shown in Figure 7.2. Modern CPLDs generally contain reprogrammable FLASH memory to hold their configuration and can programmed while in their circuits using a standard interface called "JTAG".

### 7.1.2 FPGAs

Large designs may not fit into CPLDs. Classically, such designs would have often required the development of custom chips or application-specific integrated circuits (ASICs). The ASIC market developed quite early and rapidly developed into a relatively standardized form: one could order a specific ASIC by connecting together individual standard (though ASIC-specific) elements such as gates, flip-flops, multiplexers, and memories. One did not work on the individual transistor level, but rather at more abstract level called a "gate array". In the mid 1980s, the gate array concept was extended to a user-programmable form, generating the field-programmable gate array (FPGA). The FPGA has grown in power and complexity such that a full original IBM PC could be easily implemented in a quite modest FPGA.

The main operational difference between FPGAs and CPLDs is that FPGAs do not retain their configuration in a permanent internal memory - instead it must be loaded into the FPGA at power up. This configuation step can be done from a computer over a cable (such as the JTAG cable used for programming the non-volatile configuration of the CPLD). The FPGA generally also supports loading itself from an external memory which contains


Figure 7.3: Individual logic cell of an FPGA
the "fuse list" of configuration information.
FPGAs are relatively simple in architecture. They consist of a large array of cells such as that diagrammed in Figure 7.3. The sector consists of a lookup table (memory) connected to a single flip-flop, as well as multiplexers to select the clock and output (flip-flop or combinatorial). These individual elements are then connected together by a programmable interconnect - each cell is generally directly connected to its neighbors in the array and has limited connections to cells further away. The four inputs to a cell are chosen from this set using a selection matrix - similar to an many-to-many multiplexer. At the edges of the integrated circuit, the input and output pins are connected into the interconnect matrix - they serve as the neighbors of the outer-most ring of cells. The program of the FPGA must specify both the contents of the lookup-tables and also the configuration of the interconnection matrix.

The input and output stages of the FPGAs can be quite complex. Figure 7.4 shows the "Input-Output-Buffer" (IOB) portion of the Spartan-IIe FPGA produced by Xilinx. This is a relatively old device, first available in 2001. However, the IOB is particularly clear for this device, as later devices added "double-data-rate" (DDR) capabilities which simply doubles the number of flip-flops but makes the schematic harder to understand. The IOB contains optional output and input flip-flops and also supports Tristate logic for the output. The IOB also contains programmable elements to select the logic family to be used. The Spartan IIe, an older design, supported only fifteen different logic standards. A leading-edge FPGA might support thirty. There are generally strong restrictions on the mixing of logic standards along a single edge of a device, since each logic standard may require a different reference voltage ( $V_{\text {REF }}$ ) which is shared by a number of IOBs.

While most of the silicon area of an FPGA is occupied by the logic cells and the edges are used by the IOBs, modern FPGAs have begun to add additional specialized resources. A common resource is block RAM - storage memory which is organized into rows of words in the usual way. Because the interface to the memory is somewhat constrained (compared to the generic logic cells), more memory bits can be fit into the same area of silicon. Common FPGAs may have 500 kb or more of block memory, along with 16 k one-bit registers. Other resources available in FPGAs include dedicated hardware multipliers, clock multipliers, and even microprocessors. This wide range of integrated resource allows very complex systems to be defined within a single FPGA chip.


Figure 7.4: Structure of a Spartan-IIE Input-Output Block (IOB)

### 7.2 Programmable Logic Design

The first users of PALs specifically determined the fuses to be blown by hand. With larger PALs, this task became tedious and it is effectively impossible with large modern CPLDs and FPGAs which have millions of effective fuses in their configurations. Instead, programmable logic development is performed with computer support. The computer programs are responsible for converting the developer's design, which is specified in a way which is simple for the developer, into a form appropriate for loading into the FPGA and CPLD. This process of converting the design into a "fuse-list" is called design synthesis and implementation.

Initially, the most common way technique for FPGA or CPLD development was to use the schematic interface - the developer would draw out the equivalent set of gates and flipflops and software would convert that into the values to assign to the LUTs and the interface multiplexers. Schematics are familiar to many designers and provided a fast way to develop initial simple FPGA designs. As designs became more complex, however, the schematic systems became very difficult to work with and understand. A great deal of effort was required to "lay out" the FPGA schematic neatly, even though the lay out had no effect on the actual synthesis and implementation.

The natural response to the difficulties of schematics for large projects was the development of behavioral languages. These are programming languages appropriate for describing systems of combinatorial and registered logic. Today, there are two dominant behavioral languages, often called "Hardware Design Languages" or HDLs: Verilog and VHDL. Both

Verilog and VHDL were originally developed for ASIC design and have more than enough capabilities to handle the features of all FPGAs. The choice of one language or another is generally dependent on one's training, the common language used at one's company, or personal style preferences. Stylistically, Verilog is similar to C, while VHDL is derived from the Defense Department's Ada language. Here, we will break the HDL symmetry by opting for the more concise syntax of Verilog.

### 7.2.1 Basics of Verilog

Verilog is based on a hierarchy of files or modules. Each module has a set of defined inputs and outputs. Modules can include or use other modules to form a hierarchy.

A basic Verilog module for the AND operation (as created by the Xilinx Verilog wizard) might look like this:

```
/////////////////////////////////////////////////
// Create Date: 21:44:39 01/30/2007
// Module Name: myand
// Revision:
// Revision 0.01 - File Created
// Additional Comments:
//
///////////////////////////////////////////////
module myand(a, b, y);
    input a;
    input b;
    output y;
assign y=a&b;
endmodule
```

This code describes a simple module with three ports called a, b, and y. Below the module declaration, the direction of each port is given. A port can be input, output, or inout. A signal labeled as inout can be considered both as an input and output and should be driven to high-impedance when not in use as an output (more on that later).

After the declaration of ports is the definition of the module's behavior. In this case the behavior is entirely combinatorial and consists of single assignment. An assign statement provides a combinatorial assignment. In this case, y comes directly out of an LUT with no flip flop in between.

The ports need not be single bits. They can be buses, as seen in this example:

```
module myand(a, b, y);
    input [3:0] a;
    input [3:0] b;
```

output [4:0] y;
assign $y=a+b$;
endmodule

Here $a$ and $b$ are four-bit binary numbers and $y$ is a five bit number containing their sum.

We can also operate on subsets of bits if needed, as seen here:

```
module crazy(a, b, y, z);
    input [3:0] a;
    input [3:0] b;
    output [3:0] y;
    output z;
assign y=a[2:0]+b[2:0];
assign z=a[3:1]>b[3:1];
endmodule
```


### 7.2.2 Sequential Logic

Sequential logic (including flip-flops) can also be implemented using Verilog. The syntax for sequential logic requires the definition of the clock or other control lines. The example below defines both a clocked and a combinatorial AND gate. Notice that a new input, the clock, is required for this module.

```
module dualand(a, b, y, z, clk);
    input a;
    input b;
    output y;
    output reg z;
    input clk;
assign y=a & b;
always @(posedge clk) begin
    z<=a & b;
end
endmodule
```

Notice a number of differences between y and z in this example. The port z is declared not only as an output but as a registered output : output reg. The assignment of $z$ is performed with an always loop and is performed using the "blocking assignment" $(<=)^{1}$.

### 7.2.3 Internal signals

Verilog is not limited to immediate calculations related to the input and output ports. A module can contain both combinatorial and sequential internal signals. A combinatorial internal signal is called a "wire" while a sequential internal signal is called a "reg". We can re-implement the dual-and example using internal wires and registers and leave the outputs as wires.

```
module dualand(a, b, y, z, clk);
    input a;
    input b;
    output y;
    output z;
    input clk;
    wire theand;
    reg regand;
assign theand=a & b;
assign y=theand;
assign z=regand;
always @(posedge clk) begin
    regand<=a & b;
end
endmodule
```


### 7.2.4 State Machine Example

We can very easily implement the full example scaler from Chapter 5 using Verilog. With Verilog available, it is quite easy to change the implementation from a one-hot to the morecommon binary encoded state machine. To make the code more readable, we can use an additional feature of the Verilog language: the creation of constants using parameter:

```
parameter ST_IDLE = 0;
parameter ST_COUNTING = 1;
parameter ST_OUTPUT = 2;
```

[^1]We can then implement the scaler state machine in this way:

```
module scalerState(clk, gate, reset, clear, countEnable, flag);
    input clk;
    input gate;
    input reset;
    output flag, clear,countEnable;
    reg [1:0] state;
    parameter ST_IDLE = 0;
    parameter ST_COUNTING = 1;
    parameter ST_OUTPUT = 2;
always @(posedge clk)
    case(state)
        ST_IDLE : if (gate) state<=ST_COUNTING;
            else state<=ST_IDLE;
        ST_COUNTING : if (reset) state<=ST_IDLE;
                    else if (!gate) state<=ST_OUTPUT;
                    else state<=ST_COUNTING;
        ST_OUTPUT : if (reset) state<=ST_IDLE;
                        else state<=ST_OUTPUT;
    endcase
assign clear=(state==ST_IDLE);
assign countEnable=(state==ST_COUNTING);
assign flag=(state==ST_OUTPUT);
endmodule
```


## Chapter 8

## Digital and Analog Together

The power of the modern digital computer and modern digital electronics in general is very large. Once data is in digital form, it can be moved around, transformed, combined, and generally analyzed quite easily. However, most experiments are not inherently digital: the quantities to be measured and the control signals needed to manipulate the system are actually both analog. As a result, a set of techniques is needed to convert analog voltages and currents into digital numbers and to make the opposite conversion.

### 8.1 Digital to Analog Conversion

The conversion of digital numbers into analog voltage is a relatively straightforward task. In the end, the output voltage should be proportional to the digital input. The smallest input digital value should yield the smallest voltage, while the largest digital input value should yield the largest voltage. Each increase by one in the digital input code should produce the same magnitude of increase on the analog output. Any failure of this constraint is referred to as "differential nonlinearity". The resolution, or minimum step, for a DAC can be determined from its input digital range ( $N$ bit digital number) and voltage range.

$$
R=\frac{\Delta V}{2^{N}}
$$

A DAC or ADC can only represent changes as small as the resolution or "least-count".
For some applications, it is most convenient to have an analog range from ground to a fixed positive voltage which is called a "unipolar" DAC. For other applications, a $-V_{0}$ to $V_{0}$ or "bipolar" DAC is required. Of course, either DAC can be converted to the other by application of additional op-amp circuitry.

For the purposes of discussion, we will consider various implementations of a unipolar four-bit DAC. Such a DAC has sixteen possible input values. If we assume the range of the DAC will be $0-5 \mathrm{~V}$, then each step will be $5 \mathrm{~V} / 16=0.3125 \mathrm{~V}$ and the full table will look like that below.

| DAC Input | DAC Output | DAC Input | DAC Output |
| :---: | :---: | :---: | :---: |
| 0000 | 0.0000 V | 1000 | 2.5000 V |
| 0001 | 0.3125 V | 1001 | 2.8125 V |
| 0010 | 0.6250 V | 1010 | 3.1250 V |
| 0011 | 0.9375 V | 1011 | 3.4375 V |
| 0100 | 1.2500 V | 1100 | 3.7500 V |
| 0101 | 1.5625 V | 1101 | 4.0625 V |
| 0110 | 1.8750 V | 1110 | 4.3750 V |
| 0111 | 2.1875 V | 1111 | 4.6875 V |

Notice that the full-range value of the DAC is not quite 5 V . This is typical for DACs. As the number of input bits increases (and the resolution gets finer), the maximum value will approach 5 V .

### 8.1.1 Summing DAC

Based on our experience with op-amps, we can easily create a DAC by using a summing amplifier with four inputs, one for each bit of the digital number. From Chapter 3, we recall that we can write the output of a summing amplifier in the form

$$
V=-R_{f}\left(\frac{V_{0}}{R_{0}}+\frac{V_{1}}{R_{1}}+\frac{V_{2}}{R_{2}}+\frac{V_{3}}{R_{3}}\right)
$$

If we follow the summing amplifier by a unitary-gain inverting amplifier, we can remove the sign inversion. Thus, the key question is the choice of the resistors.

From the DAC table, we see that a change in the lowest bit of the digital number will produce a change in the DAC output of 0.3125 V , while a change in the second-lowest bit will produce a change in DAC output of 0.6250 V : a factor of two larger. This is exactly the behavior one would expect from binary numbering. Thus, we want $R_{0}=2 R_{1}$. In fact, we want each bit's resistor to vary by a factor of two from the next. We can choose one resistor as the "standard" and scale all the rest to that resistor - the simplest choice is to pick the resistor for the most-significant bit.

$$
\begin{array}{cc}
R_{3}=R & R_{2}=2 R \\
R_{1}=4 R & R_{0}=8 R
\end{array}
$$

We must then choose $R_{f}$ to provide the proper total dynamic range. When all bits are on, we find

$$
\frac{15}{16} V=R_{f}\left(\frac{V_{+}}{8 R}+\frac{V_{+}}{4 R}+\frac{V_{+}}{2 R}+\frac{V_{+}}{R}\right)=V_{+} \frac{15}{8} \frac{R_{f}}{R}
$$

If we assume the digital inputs are $0-5 \mathrm{~V}$ (e.g. 74 HC TTL), then we can determine that $R_{f}=\frac{1}{2} R_{f}$. If the digital inputs have a different range, we would need a different feedback resistor. The resulting DAC is shown in Figure 8.1.

While such a DAC is easy to design, based on our experience with analog electronics, it is relatively difficult to construct a accurate summing DAC, particularly one with a large number of bits. A large number of resistors with very precise values are needed - the


Figure 8.1: A digital-to-analog converter based on a summing amplifier configuration.
precision of the ratios determines the differential non-linearity of the device. It is difficult to accurately construct a full range of resistors with relative values from 1 to 2048, as would be needed for a 12 -bit DAC.

### 8.1.2 R-2R Ladder



Figure 8.2: A digital-to-analog converter based on an R-2R ladder summing amplifier configuration.

The difficulty of accurately constructing a wide range of resistor values can be bypassed through the use of a current-dividing resistor network called an "R-2R Ladder". In an R-2R ladder, the "rungs" of the ladder are 2 R resistors while the vertical portion of the ladder is


Figure 8.3: Pulse-train output for a 4-bit PWM for a range of different input digital values.
a series of resistors with resistance $R$. The structure of a voltage DAC constructed using an R-2R ladder is shown in Figure 8.2. The various input digital bits control switches which either connect the far end of the 2 R resistor directly to ground or to the a summing amplifier's virtual ground. In either case, the effective voltage on the lower end of each 2 R resistor is ground. The $\mathrm{R}-2 \mathrm{R}$ ladder requires the creation of just two resistor values with a simple relationship between them. The circuit is relatively insensitive to the absolute magnitude of R and will be stable even with temperature changes as long as the $\mathrm{R} / 2 \mathrm{R}$ ratio remains stable.

To understand the operation of the $\mathrm{R}-2 \mathrm{R}$ ladder, it is important to notice that the resistance to ground on the right of any 2 R resistor is always 2 R . As a result, the ladder behaves as a uniform current divider. At the first 2 R resistor, half the current flows through the 2 R resistor and half continues down the chain. At the next 2 R resistor, half the remaining current (a quarter of the total) flows through the 2 R resistor and the rest continues down the chain. This process can continue for as many bits as necessary - each 2 R resistor will pass half the current of the one to its left. When the digital switch connects the base of the 2 R resistor to the summing amplifier, the current flows through the summing amplifier and produces a voltage signal on the output. By selecting a feedback resistance of R , the DAC will properly provide a voltage of 2.5 V in the 1000 case as expected. The simplicity of the resistor arrangement makes the R-2R structure one of the most commonly used. Often, the R-2R ladder is provided without the op-amp, which allow the DAC to be used a current-mode DAC or voltage mode by adding an external op-amp.

### 8.1.3 Pulse-Width Modulation

Digital to analog conversion can also be performed with time-dependent techniques which are often much simpler than the full analog constructs. One of the simplest techniques to implement is pulse-width modulation (PWM). In pulse-width modulation, the DAC outputs a series of pulses with fixed frequency but a duty-cycle which depends on the digital input as seen in Figure 8.3. When the digital input is close to zero, the output pulses are high for a small fraction of the pulse period. When the digital input is close to maximum, the output pulses are high for nearly the entire period. The output of the PWM must then be low-pass filtered. The characteristic frequency of the low-pass filter must be low enough remove the
transients associated with the digital edges of the PWM signal. As a rule of thumb, one would require

$$
\omega_{R C}=\frac{1}{R C}<0.01 \frac{2 \pi}{T_{P W M}}
$$

The final output of the filter is a analog signal proportional to the input digital value. When PWM is used for a time-varying input, the PWM period must be much shorter than the time variation for the digital inputs.

The digital portion of a PWM can be very easily implemented in programmable logic. The Verilog module below provides a complete implementation for a four-bit PWM DAC.

```
module pwm(clk, digitalIn, analogOut);
    input clk;
    input [3:0] digitalIn;
    output analogOut;
    reg [3:0] counter;
always @(posedge clk)
    counter<=counter+1;
assign analogOut=(counter<digitalIn);
endmodule
```

The resolution of the DAC can be increased by simply increasing the number of bits in the input and the counter. However, this change will increase the effective period of the PWM since

$$
T_{P W M}=\frac{2^{N}}{f_{c l k}}
$$

Thus, an increase in the number of bits will require a lower low-pass filter cutoff or a higher clock frequency for the PWM. The low-pass filter behavior will also affect desired output changes (for example when the input digital value changes) while all devices have maximum effective clock frequencies.

Despite its simplicity, the PWM DAC is a very effective DAC. In particular, the differential non-linearity of a PWM DAC is very small - there is no need to carefully match resistor values to guarantee that all steps in the DAC's range are of the same size. With the increasing availability of high-speed digital electronics, PWM techniques are seeing have found wide application in a range of analog output applications.

### 8.1.4 Sigma-Delta Conversion

The PWM technique is quite effective, but has a few flaws. One of the most significant is that the fixed period of the rising edges of the PWM generates overtones which may distort

## Special Topic: PWM for high power applications

Many experiments require control of high-power or high-current devices such as resistive heaters or powerful motors. Often it is insufficient to simply have such heater or motor in the on or off state, but rather it is desirable to have more continuous control over its operation. For example, in the case of a heater one might want a feedback loop to provide a constant heating power, rather than flipping the heater on and off with the resulting large heating and cooling cycles. Similar requirements apply to electric motors, high current lights, and other devices.

Traditionally, such control has been provided by linear transistor-based devices where the current provided is proportional to a voltage input signal, as in the figure below. Such devices behave like variable resistors whose resistance can change from zero to very large and thus control the current flowing through the device connected in series with them. The power dissipated in the variable resistor is

$$
P=I^{2} R
$$

so that no power is dissipated in the device when the effective resistance is either zero or very large - when the resistance is large, the current will be zero. However, when the resistance is in between, very large amounts of power can be dissipated in the device. As a result, traditional linear controllers are often very massive with heavy heat-transport needs.


More recently, traditional linear controllers are falling out of favor compared to PWM controllers. A complete integrated PWM controller has an analog input and an internal analog-to-digital converter simply to drive the PWM "DAC" output. The heater or electric motor is generally able to perform the low-pass filter function directly, without any need for capacitors. A control transistor is still required, but that transistor is now always either fully open or fully closed - and no power is being dissipated in the transistor. The result is a more power-efficient design and a longer life for the controller.
an analog waveform or inject high-frequency noise into the system. To combat these effects, very short pulse periods (high frequencies) must be used in combination with strong lowpass filtering as discussed above. These effects are large for the PWM because the periodic component is at a very low frequency - the fundamental $T_{P W M}$.

Alternatively, one can keep the width of the pulse fixed and simply change the frequency of pulses. There are several ways to achieve this, but one of the most effective is to calculate a "running error" on the output compared to the input and use this to determine when the output should change value. This technique is called "sigma-delta" conversion. The sigmadelta technique uses a single output bit, just as for PWM. This output has two possible states, either zero or the maximum output voltage. The converter sums the requested output with the recent history and the current output to determine the next output.

For a DAC with a number of bits $N$ and an input value of $p$, the summing process as a function of clock cycle $k$ is

$$
\begin{aligned}
s_{k+1} & =p_{k+1}+s_{k}+ \begin{cases}0, & V_{k}=0 \\
-2^{N}, & V_{k}=V_{r e f}\end{cases} \\
V_{k+1} & = \begin{cases}0, & s_{k+1}<2^{N} \\
V_{r e f}, & s_{k+1} \geq 2^{N}\end{cases}
\end{aligned}
$$

The behavior of a four-bit sigma-delta converter can be seen in Figure 8.4. With a sigmadelta converter, the converter will have a fairly steady "overtone" frequency for any given input code, but the overtone frequencies will be different for different input codes. Thus, for a rapidly modulated signal (such as an audio signal), no particular harmonic will dominate. The sigma-delta also immediately adjusts to a changed input code - the summing cycle will take into the account the changed value in the next iteration, rather than waiting for the full $T_{P W M}$ as in the PWM case. These advantages make the converter ideal for audio applications.

This converter can be represented in Verilog quite simply in the following way:

```
module sigmaDelta(clk, digitalIn, analogOut);
    input clk;
    input [7:0] digitalIn;
    output analogOut;
    reg [8:0] accumulator;
always @(posedge clk)
    if (analogOut) accumulator<=accumulator+digitalIn-9'h100;
    else accumulator<=accumulator+digitalIn;
assign analogOut=accumulator[8];
endmodule
```



Figure 8.4: Accumulator sum and output pulse train for a 4-bit sigma-delta DAC for two different input digital values (4 and 14).

### 8.2 Analog to Digital Conversion

In the same way that there are many ways to convert digital into analog, there are also a number of ways to convert analog into digital. Each of the various techniques has strengths and weaknesses in the various key ADC parameters. The most obvious ADC parameter is the resolution of the ADC - the smallest resolvable difference between voltages. For a usual linear ADC, the resolution has the same form as a linear DAC:

$$
R=\frac{\Delta V}{2^{N}}
$$

Thus, a larger number of bits is generally better. However, ADCs must also be characterized by the time required for a conversion from analog to digital. ADCs are not arbitrary fast and a larger number of bits generally implies a longer conversion time. Finally, the noise sensitivity of the ADC may be important, depending on the application.

### 8.2.1 ADC Helper: the Sample-and-Hold

For many ADC implementations, the actual ADC is preceded by a sample-and-hold circuit. This is an analog circuit designed to keep the input to the ADC logic stable during a conversion. An example of a sample-and-hold circuit is shown in Figure 8.5. In this circuit, the capacitor is responsible for holding the voltage constant during the ADC conversion
process. The buffer amplifier after the capacitor is responsible for minimizing the droop of the capacitor during the conversion process by providing a high-impedance input, while the buffer amplifier before the capacitor is responsible for rapidly changing the capacitor's charge when the switch is closed. The switch itself would be a CMOS gate under the control of the ADC master control logic.


Figure 8.5: Example of a sample-and-hold circuit.

### 8.2.2 Flash ADC

The fastest form of ADC is called a Flash-ADC. A flash-ADC contains a voltage divider chain and a large number of comparators ( $2^{n}-1$ ). The voltage divider chain produces the $2^{n}$ voltages associated with each individual code transition. The input voltage is compared to each of these voltages, producing $2^{n}-1$ comparisons.

$$
C_{k}=V_{i n}>\left(V_{r e f} \frac{k R}{2^{n} R}\right)
$$

The output of these comparisons indicates the voltages which are just smaller and just larger than the input voltage: the $k$ when the $C_{k}=0$ for the first time is the transition point. The comparator outputs can then be fed into a priority encoder to determine the digital number associated with $k$. A simple example of such an ADC can be seen in Figure 8.6.

Flash-ADCs are very fast, since they depend only on combinatorial logic. However, they are quite complicated - particularly for large number of bits. Since the number of resistors and comparators scales with $2^{n}$, flash-ADCs are generally restricted to relatively small numbers of bits - generally eight bits or less. Since a flash-ADC makes many "useless" comparisons, flash-ADCs tend to require a large amount of electrical power as well, compared to other technologies.

### 8.2.3 Integration ADC

On the opposite side of the spectrum from the Flash-ADCs in terms of resolution and speed is the integration ADC. Integration ADCs operate by determining the amount of time necessary to integrate a given amount of charge. Consider the circuit in Figure 8.7 which is the core of a dual slope integrator.


Figure 8.6: A two-bit flash-ADC.


Figure 8.7: The core integrator and comparator for a dual-slope ADC.

The dual slope integrator works by connecting the input voltage to an op-amp integrator for a fixed period of time $T_{i n}$. This time is generally set in terms of clock cycles on an internal oscillator. During this first integration period, a current

$$
I_{i n}=\frac{V_{i n}}{R}
$$

will flow into the capacitor, causing it to charge to a total voltage

$$
V_{\max }=-\frac{Q_{i n}}{C}=-\frac{V_{i n} T_{i n}}{R C}
$$

The dual-slope integrator then changes the input of the integrator a reference voltage $-V_{\text {ref }}$. This causes the integrator to slowly discharge with a constant current

$$
I_{r e f}=\frac{-V_{r e f}}{R}
$$

The ADC then determines the amount of time required for the output of the integrator to fall back to ground : $T_{\text {ref }}$. Once the integrator has returned to ground, an equal amount of charge must have been removed compared to what was added in the initial integration step.

$$
\begin{aligned}
Q_{i n} & =Q_{r e f} \\
I_{i n} T_{i n} & =-I_{r e f} T_{r e f} \\
\frac{V_{i n}}{R} T_{i n} & =\frac{V_{r e f}}{R} T_{r e f} \\
V_{i n} & =V_{r e f} \frac{T_{r e f}}{T_{i n}}
\end{aligned}
$$

Thus, the input voltage is completely determined by the reference voltage and the two times, which can be measured by digital counters of whatever precision is required. Notice that the values of the capacitor and resistor do not enter into the final result, so the measurement is insensitive to drifts in the capacitor and resistor values due to temperature or aging effects - as long as the change during the integration cycle is small.

The stability and precision of dual-slope integration ADCs motivates their use in test equipment. Most digital multimeters (DMMs) operate on the dual-slope principle as it allows for very high resolution. The integration technique also greatly reduces the noise sensitivity of the device - all high-frequency noise is averaged out leaving a very stable voltage reading. The major drawback of the dual-slope technique is that it is generally very slow - a few Hertz repetition rate is generally the limit for the technique.

### 8.2.4 Successive-Approximation ADC

Between the flash-ADC and the integrator ADC , there is clearly a need for a moderatecomplexity but reasonably fast ADC. This niche is generally filled by one of several ADCs which depend on the presence of a digital-to-analog converter (DAC) and an analog comparator to determine the relationship between the DAC output and the input voltage.

The simplest sort of ADC based on a DAC would simply connect the DAC to a counter. The counter would begin at zero (lowest DAC output voltage) and count upwards as long as the comparator indicated that the input voltage was bigger than the DAC output. When the comparator switched, the counter would stop and the value on the counter would be the ADC output. This technique is very simple, but rather slow and non-deterministic - the ADC conversion time depends on the input voltage. If the input voltage is low, the ADC converts quickly, while if the input voltage is high, the counter must count longer and the ADC is slower. This sort of behavior is generally not acceptable when acquiring data to be used for Fourier analysis or other time-dependent techniques. These techniques require samples taken at a fixed frequency.

The successive-approximation technique allows a deterministic ADC conversion requiring only $N+1$ clocks or DAC output values, where $N$ is the number of bits in the ADC digital output. By comparison, a counter-ADC would require on average $2^{N-1}$ clocks. The successive-approximation ADC operates by beginning setting the most-significant bit (MSB) feeding the DAC to high and all the rest low. This will cause the DAC to output its median output voltage. At the next clock edge, the final value of the MSB is determined


Figure 8.8: DAC output voltages during a four-bit successive approximation process. At each clock, a comparator determines if the input value is above or below the DAC value, determining the DAC value for the next step.
by the output of the comparator. If the input voltage is greater than the DAC value, the bit will be kept high, while if the input voltage is lower than the DAC value, then the bit will be set low. The MSB is then kept fixed and the next bit in line is set high and the process repeats. After each clock cycle, the value of another bit is determined until the value of the least-significant bit finally determined and the process is complete. The successiveapproximation process can be represented by a "decision tree" as shown in Figure 8.8 which shows all the possible DAC values in the successive-approximation process and the possible next steps from each DAC output value. Because of their speed and relative compactness, successive-approximation ADCs are fairly widely used.

The key to the successive-approximation ADC is the "successive-approximation register"(SAR) which receives the input from the comparator and determines the value to be sent to the DAC. While an SAR is often simply part of an integrated ADC, we can define one in Verilog for use with an external DAC or even with a PWM.

```
module sar(clk, start, compare, digitalOut, done);
    input clk, start, compare;
    output reg [3:0] digitalOut;
    output done;
    reg [2:0] counter;
always @(posedge clk) begin
    if (start && counter==0) begin
```

```
    digitalOut[3]<=1;
    digitalOut[2:0]<=0;
    counter<=counter+1;
    end else if (counter==1) begin
    digitalOut[3]<=compare;
    digitalOut[2]<=1;
    counter<=counter+1;
    end else if (counter==2) begin
    digitalOut[2]<=compare;
    digitalOut[1]<=1;
    counter<=counter+1;
    end else if (counter==3) begin
    digitalOut[1]<=compare;
    digitalOut[0]<=1;
    counter<=counter+1;
    end else if (counter==4) begin
    digitalOut[0]<=compare;
    counter<=counter+1;
    end else if (counter==5) counter<=0;
end
assign done=(counter==5);
endmodule
```


### 8.3 Time to Digital Conversion

Conversion from analog voltage (or current) to digital value is important, but sometimes it is also necessary to perform what is sometimes called "time to digital" conversion (TDC). The inputs for time-to-digital conversion are actually digital pulses. The conversion in question is the determination of the time between digital pulses - either pulses on the same wire or between a pulse on one wire and pulse on another. This technique is needed for very precise measurements of speed as well as for determine the relative ordering of events within an experiment. For a particle-beam experiment, time-to-digital conversion is often useful to determine (for example) particle mass by comparing the momentum of a particle to its velocity, determined by the time required to cover a fixed distance.

### 8.3.1 Pure Digital Techniques

For time periods which are within the range of conventional digital electronics, a TDC can be constructed using a simple counter. One signal starts the counter and the other signal stops the counter - the counter's clock determines the resolution of the TDC. This technique is conceptually similar to a chess clock which is started by the hand of one participant and


Figure 8.9: Current-integrating analog TDC
stopped by the other. A more sophisticated technique does not actually stop the counter when the second pulse arrives, but merely stores the value of the counter at that moment into a memory. This allows the TDC to also count possible subsequent pulses - their times can be added to the memory after the first. Such techniques can be easily implemented in FPGA devices and operate easily up to 400 MHz ( 2.5 ns ) in commercial devices. Custom digital TDCs are able to operate with speeds above 1 GHz , with resolutions of 500 ns or so.

### 8.3.2 Analog-based Techniques

For very short time periods, digital techniques are not sufficient: most counter designs in the most recent silicon processes reach their limits at around 300 ps . For shorter periods, one can use an analog technique which is essentially similar to the slope-integration ADC technique. A precision analog TDC, as pictured in Figure 8.9, contains a precision constant-current source which can be connected either to ground or to the virtual ground of an integrator circuit. The integrator is initially discharged and the current source is disconnected from the integrator. The first pulse connects the current source to the integrator and the second pulse disconnects the current source again. The amount of integrated charge is proportional to the time that the current source was connected to the integrator.

$$
\begin{aligned}
Q & =C V \\
I \Delta T & =C V \\
\Delta T & =\frac{C}{I} V
\end{aligned}
$$

Thus, the integrator is followed by a conventional voltage ADC (such as a successiveapproximation ADC) to determine the voltage. The voltage can be combined with the known capacitance and current to find the time.

The analog current-integration technique depends on fast analog switches to make and break the connections to the integrator. With the fast switches, however, such designs are capable of operating with resolutions as small as 35 ps . Since even light travels only 1 cm in 35 ps , such a TDC requires very carefully-managed inputs to achieve its full performance.

Most sensors will produce variations at a much larger scale than 100 ps , even for events occurring at precisely the same time. Such timing variation for nominally synchronous events is referred to as jitter. Jitter is a form of noise in timing measurements.


[^0]:    ${ }^{1}$ While the impedance of op-amp input can be taken as infinite, it does not follow that the input impedance of all op-amp circuits will be infinite!

[^1]:    ${ }^{1}$ Technically, the regular assignment can be used but frequently leads to confusion. For almost all purposes, the blocking assignment should be used.

