Topics on Electrical Engineering

D. T. McGuiness, Ph.D

M.Sc Non-Linear Electronics

Lecture Book

2025.WS

Table of Contents

Part I Circuit Fundamentals

Chapter 1 **Operational Amplifiers** **3**

1.1 Introduction 3
1.2 The Ideal Op Amp 5
Terminals of the Element · Function and Characteristics · Differential and Common-Mode Signals
1.3 The Inverting Amplifier 9
The Closed-Loop Gain · Effect of Finite Open-Loop Gain · Input and Output Resistances
1.4 Non-Inverting Amplifier 14
The Closed-Loop Gain · Effect of Finite Open-Loop Gain · Input and Output Resistance · The Voltage Follower
1.5 Difference Amplifiers 17
A Single-Op-Amp Difference Amplifier · A Versatile Circuit: The Instrumentation Amplifier
1.6 Integrators and Differentiators 24
The Inverting Configuration with General Impedances · The Inverting Integrator · The Op-Amp Differentiator
1.7 DC Imperfections 29
Offset Voltage · Input Bias and Offset Currents · Effects of Offset of Inverting Integrator
1.8 Finite Open-Loop Gain and Bandwidth 34
Frequency Dependence of the Open-Loop Gain · Frequency Response of Closed-Loop Amplifiers
1.9 Large-Signal Operation of Op Amps 37
Output Voltage Saturation · Output Current Limits · Slew Rate · Full-Power Bandwidth

Chapter 2 **Semiconductors** **41**

2.1 Introduction . . . 41
2.2 Intrinsic Semiconductors . . . 43
2.3 Doped Semiconductors . . . 46
2.4 Current Flow in Semiconductors . . . 48
Drift Current · Diffusion Current
2.5 The PN Junction . . . 52
Implementation · Open Circuit Operation
2.6 PN Junction with Applied Voltage . . . 57
Qualitative Description of Junction Operation · Reverse Breakdown
2.7 Capacitive Effects in the pn Junction . . . 61
Depletion or Junction Capacitance · Diffusion Capacitance

Part II Semiconductor Devices

Supplemental List of Items

List of Figures

1.1 On of the first commercially available op-amps, the $\mu A709$. 3
1.2 The traitorous eight. From left to right: Gordon Moore, C. Sheldon Roberts, Eugene Kleiner, Robert Noyce, Victor Grinich, Julius Blank, Jean Hoerni and Jay Last (1960). 4
1.3 Circuit symbol for the standard op-amp. 5
1.4 The op amp shown connected to dc power supplies. 5
1.5 Equivalent circuit of the ideal op amp. 6
1.6 Representation of the signal sources v_1 and v_2 in terms of their differential and common-mode components. 8
1.7 The inverting amplifier in closed-loop configuration. 9
1.8 Analysis of the inverting configuration taking into account the finite open-loop gain of the op amp. 11
1.9 The noninverting configuration. 14
1.10 (a) The unity-gain buffer or follower amplifier. (b) Its equivalent circuit model. . . 16
1.11 A difference amplifier. 18
1.12 Application of superposition to the analysis of the circuit 18
1.13 Finding the input resis-tance of the difference amplifier for thecase R3= R1 and R4= R2. 20
1.14 A popular circuit for an instrumentation amplifier. (a) Initial approach to the circuit. (b) The circuit in (a) with the connection between node X and ground removed and the two resistors R1 and R1 lumped together. This simple wiring change dramatically improves performance. 21
1.15 The inverting configuration with general impedances in the feedback and the feed-in paths. 24
1.16 Frequency response of the integrator. 26
1.17 The Miller integrator with a large resistance RF connected in parallel with C in order to provide negative feedback and hence finite gain at dc. 26
1.18 Circuit model for an op amp with input offset voltage V_{OS}. 29
1.19 The output dc offset voltage of an op amp can be trimmed to zero by connecting a potentiometer to the two offset-nulling terminals. The wiper of the potentiometer is connected to the negative supply of the op amp. 30
1.20 (a) The op-amp input bias currents represented by two current sources IB1 and IB2 (b) Analysis of the closed-loop amplifier, taking into account the input bias currents. 31

1.21 (a)In an ac-coupled amplifier the dc resistance seen by the inverting terminal is R2 ; hence R3 is chosen equal to R2. (b) Illustrating the need for a continuous dc path for each of the op-amp input terminals. Specifically, note that the amplifier will not work without resistor R3. 32

2.1 The periodic table of elements. 43
2.2 A bar of silicon (a) into which holes are injected, thus creating the hole concentration profile along the x axis, shown in (b). The holes diffuse in the positive direction of x and give rise to a hole diffusion current in the same direction. Note that we are not showing the circuit to which the silicon bar is connected. 50
2.3 If the electron concentration profile shown is established in a bar of silicon, electrons diffuse in the x direction, giving rise to an electron diffusion current in the negative-x direction. 50
2.4 Simplified physical structure of the pn junction. As the pn junction implements the junction diode, its terminals are labeled anode and cathode. 52
2.5 (a) The pn junction with no applied voltage (open-circuited terminals). (b) The potential distribution along an axis perpendicular to the junction. 53
2.6 The electron density of a MESFET. Please observe the central point where the environment is void of electrons. 53
2.7 57

List of Tables

List of Examples

List of Theorems

Part I

Circuit Fundamentals

Part Contents

Chapter 1 Operational Amplifiers 3

Chapter 2 Semiconductors 41

It has today occurred to me that an amplifier using semiconductors rather than vacuum is in principle possible.

(William Shockley, Laboratory notebook, 29 Dec 1939.)

Chapter 1

Operational Amplifiers

Table of Contents

1.1 Introduction 3
1.2 The Ideal Op Amp 5
1.3 The Inverting Amplifier 9
1.4 Non-Inverting Amplifier 14
1.5 Difference Amplifiers 17
1.6 Integrators and Differentiators 24
1.7 DC Imperfections 29
1.8 Finite Open-Loop Gain and Bandwidth 34
1.9 Large-Signal Operation of Op Amps 37

1.1 Introduction

Having learned basic amplifier concepts and terminology in our previous chapter, we now have the necessary tools to study a circuit building block of universal importance:

> The operational amplifier, or simply **op-amp**.

Figure 1.1
On of the first commercially available op-amps, the $\mu A709$.

Op amps have been in use for a long time, with their initial applications being primarily in the areas of analog computation and sophisticated instrumentation. Early op-amps were constructed from discrete components[1], and their cost comparatively high. In the mid 1960s is the first Integrated Circuit (IC) op-amp was produced. This was made up of a relatively large number of transistors and resistors all on the same silicon chip. Although its characteristics were poor[2] and its price was still quite high, its appearance signaled a new era in electronic circuit design.

[1] vacuum tubes and then transistors, and resistors

[2] by today's standards.

Engineers started using op amps in large quantities, which caused their price to significantly drop. They also demanded better-quality op amps. Semiconductor manufacturers responded quickly, and within the span of a few years, high-quality op amps became available at extremely low prices[3] from a large number of suppliers.

[3] It is more than possible to buy a dozen of IC op-amps for less than a euro.

One of the reasons for the popularity of the op amp is its **versatility**. As we will shortly see in this chapter, we can do almost anything with op amps. Equally important is the fact that the IC op amp has characteristics that closely approach the assumed ideal. This implies that it is quite easy to design circuits using the IC op amp. Also, op-amp circuits work at performance levels that are quite close to those predicted theoretically. It is for this reason that we are studying op amps at this early stage.

As already implied, an IC op amp is made up of a large number[4] of transistors together with resistors, and one capacitor connected in a rather complex circuit.[5]

[4] about 20

> Given we have **NOT** yet studied transistor circuits, the circuit inside the op amp will not be discussed in this chapter.

Rather, we will treat the op amp as a circuit building block and study its terminal characteristics and its applications. This approach is quite satisfactory in many op-amp applications.

[5] An example of an IC op-amp, LM358 Dual Op-Amp IC.

Information: The Traitorous Eight

The *traitorous eight* was a group of eight (**8**) employees who left **Shockley Semiconductor Laboratory** in 1957 to found **Fairchild Semiconductor**. William Shockley had in 1956 recruited a group of young Ph.D. graduates with the goal to develop and produce new semiconductor devices. While Shockley had received a Nobel Prize in Physics and was an experienced researcher and teacher, his management of the group was authoritarian and unpopular. This was accentuated by Shockley's research focus not proving fruitful. After the demand for Shockley to be replaced was rebuffed, the eight left to form their own company.

Figure 1.2
The traitorous eight. From left to right: Gordon Moore, C. Sheldon Roberts, Eugene Kleiner, Robert Noyce, Victor Grinich, Julius Blank, Jean Hoerni and Jay Last (1960).

Shockley described their leaving as a "betrayal". The eight who left Shockley Semiconductor were Julius Blank, Victor Grinich, Jean Hoerni, Eugene Kleiner, Jay Last, Gordon Moore, Robert Noyce, and Sheldon Roberts. In August 1957, they reached an agreement with Sherman Fairchild, and on September 18, 1957, they formed Fairchild Semiconductor. The newly founded Fairchild Semiconductor soon grew into a leader in the semiconductor industry.

In 1960, it became an incubator of Silicon Valley and was directly or indirectly involved in the creation of dozens of corporations, including **Intel** and **AMD**. These many spin-off companies came to be known as "Fairchildren".

1.2 The Ideal Op Amp

1.2.1 Terminals of the Element

From a signal point of view the op-amp has three (3) terminal

two (2) input terminals and one (1) output terminal.

Fig. 1.3 shows the symbol we shall use to represent the op amp. Terminals 1 and 2 are input terminals, and terminal 3 is the output terminal. As explained in the previous Chapter, amplifiers require Direct Current (DC) power to operate.[6] Most IC op-amps require two (2) DC power supplies, as shown in **Fig.** 1.4. Two terminals, 4 and 5, are brought out of the op-amp package and connected to a positive voltage V_{CC} and a negative voltage $-V_{EE}$, respectively. In Fig. 2 (*b*) we explicitly show the two dc power supplies as batteries with a common ground. It is interesting to note that the reference grounding point in op-amp circuits is just the common terminal of the two power supplies; that is, no terminal of the op-amp package is physically connected to ground. In what follows we will not, for simplicity, explicitly show the op-amp power supplies.

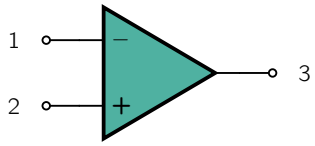


Figure 1.3
Circuit symbol for the standard op-amp.

[6] The power to amplify a given signal must come from somewhere.

> In addition to the three signal terminals and the two power-supply terminals, an op amp may have other terminals for specific purposes. These other terminals can include terminals for frequency compensation and terminals for offset pulling.

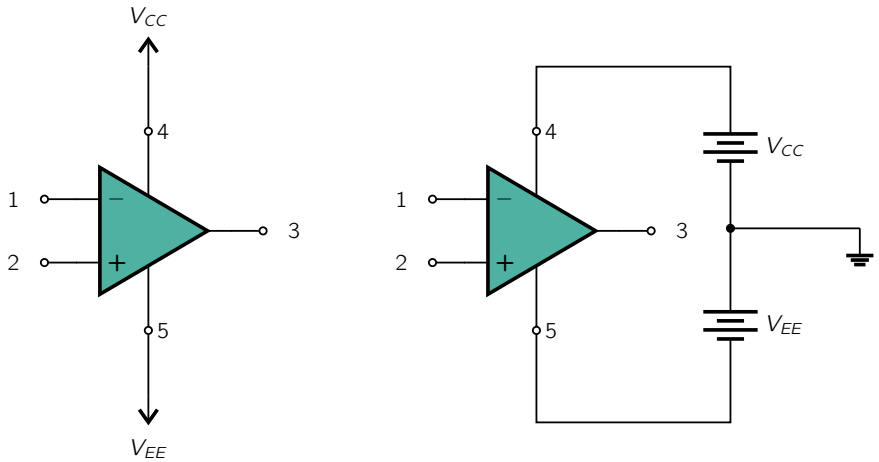


Figure 1.4
The op amp shown connected to dc power supplies.

1.2.2 Function and Characteristics

We now consider the circuit function of the op amp. The op amp is designed to sense the difference between the voltage signals applied at its two (2) input terminals,[7] multiply this by a number A, and cause the resulting voltage $A(v_2 - v_1)$ to appear at output terminal 3. Thus $v_3 = A(v_2 - v_1)$. Here it should be emphasized that when we talk about the voltage at a terminal we mean the voltage between that terminal and ground; thus v_1 means the voltage applied between terminal 1 and ground.

[7] i.e., the quantity $v_2 - v_1$

> The ideal op amp is **NOT** supposed to draw any input current.

that is, the signal current into terminal 1 and the signal current into terminal 2 are both zero. In other words, the input impedance of an ideal op amp is supposed to be infinite.

How about the output terminal 3?

This terminal is supposed to act as the output terminal of an ideal voltage source. That is, the voltage between terminal 3 and ground will always be equal to $A(v_2 - v_1)$, independent of the current that may be drawn from terminal 3 into a load impedance.

> In other words, the output impedance of an ideal op amp is supposed to be zero.

Putting together all of the above, we arrive at the equivalent circuit model shown in **Fig.** 1.5. Note that the output is in phase with (as the same sign as) v_2 and is out of phase with (has the opposite sign of) v_1.

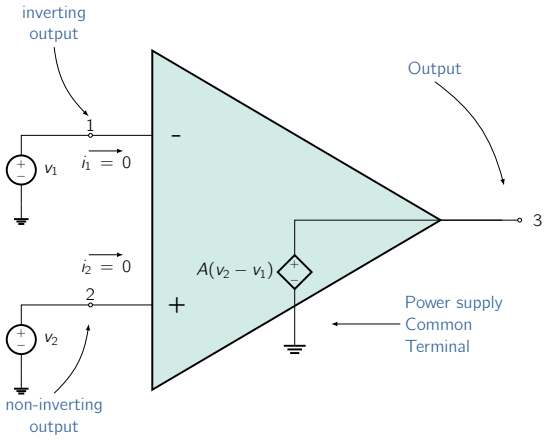


Figure 1.5 Equivalent circuit of the ideal op amp.

For this reason:

- input terminal 1 is called the **inverting input terminal** and is distinguished by a "−" sign,
- while input terminal 2 is called the **non-inverting input terminal** and is distinguished by a "+" sign.

As can be seen from **Fig.** 1.5, the op amp responds only to the *difference* signal $v_2 - v_1$ and hence ignores any signal common to both inputs. That is, if $v_1 = v_2 = 1$ V, then the output will (ideally) be zero. We call this property **common-mode rejection**, and we conclude an ideal op amp has zero common-mode gain or, equivalently, infinite common-mode rejection. We will have more to say about this point later. For the time being note that the op amp is a **differential-input**, **single-ended-output** amplifier, with the latter term referring to the fact that the output appears between terminal 3 and ground.

Furthermore, gain A is called the **differential gain**, for obvious reasons. Perhaps **NOT** so obvious is another name that we will attach to A:

the open-loop gain.

The reason for this name will become obvious later on when we "close the loop" around the op amp and define another gain, the closed-loop gain.

An important characteristic of op amps is that they are direct-coupled or dc amplifiers, where dc stands for direct-coupled.[8] The fact that op amps are direct-coupled devices will allow us to use them in many important applications.

[8] it could equally well stand for direct current, since a direct-coupled amplifier is one that amplifies signals whose frequency is as low as zero.

> Unfortunately, though, the direct-coupling property can cause some serious practical problems, as will be discussed in a later section.

How about bandwidth? The ideal op amp has a gain A which remains constant down to zero frequency and up to infinite frequency. That is, ideal op amps will amplify signals of any frequency with equal gain, and are thus said to have infinite bandwidth.

We have discussed all of the properties of the ideal op amp except for one, which in fact is the most important. This has to do with the value of A. The ideal op amp should have a gain A whose value is very large and ideally infinite. One may justifiably ask:

If the gain A is infinite, how are we going to use the op amp?

The answer is very simple:

In almost all applications the op amp will **NOT** be used alone in a so-called open-loop configuration. Rather, we will use other components to apply feedback to close the loop around the op amp. For future reference, lists the characteristics of the ideal op amp.

- Infinite input impedance
- Zero output impedance
- Zero common-mode gain or, equivalently, infinite common-mode rejection
- Infinite open-loop gain A
- Infinite bandwidth

1.2.3 Differential and Common-Mode Signals

The differential input signal v_{Id} is simply the **difference** between the two (2) given input signals v_1 and v_2, which we can write it as:

$$v_{1d} = v_2 - v_1 \tag{1.1}$$

The common-mode input signal v_{Im} is the **average** of the two (2) input signals v_1 and v_2; namely,

$$v_{Icm} = \frac{1}{2}(v_1 + v_2) \tag{1.2}$$

Both Eq. (**1.1**) and Eq. (**1.2**) can be used to express the input signals v_1 and v_2 in terms of their differential and common-mode components as: follows:

$$v_1 = v_{Im} - v_{Id}/2 \tag{1.3}$$

and

$$v_2 = v_{Im} + v_{Id}/2 \tag{1.4}$$

These equations can in turn lead to the pictorial representation in **Fig.** 1.6.

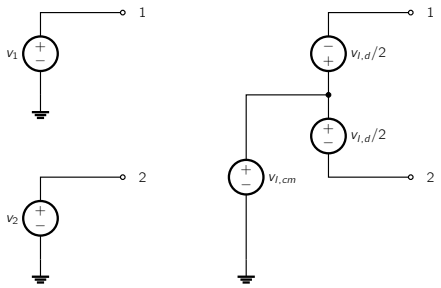


Figure 1.6: Representation of the signal sources v_1 and v_2 in terms of their differential and common-mode components.

1.3 The Inverting Amplifier

As mentioned previously, op amps are not used alone; rather, the op amp is connected to passive components in a feedback circuit. There are two (2) such basic circuit configurations employing an op amp and two (2) resistors:

1. inverting configuration,
2. non-inverting configuration.

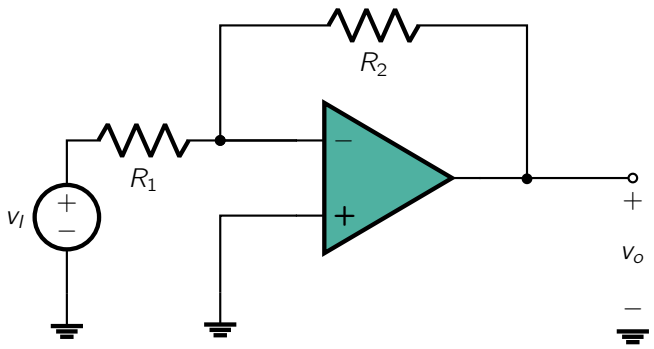


Figure 1.7
The inverting amplifier in closed-loop configuration.

Fig. 1.7 shows the inverting configuration. It consists of one op amp and two (2) resistors:

R_1 and R_2

Resistor R_2 is connected from the output terminal of the op amp (terminal 3) back to the inverting or negative input terminal (terminal 1). We speak of R_2 as applying **negative feedback**; if R_2 were connected between terminals 3 and 2 we would have called this **positive feedback**.

> Please observe that R_2 also closes the loop around the op amp.

In addition to adding R_2, we have grounded terminal 2 and connected a resistor R_1 between terminal 1 and an input signal source with a voltage v_t. The output of the overall circuit is taken at terminal 3.[9] Terminal 3 is, of course, a convenient point from which to take the output, since the impedance level there is ideally zero. Therefor, the voltage v_0 will not depend on the value of the current that might be supplied to a load impedance connected between terminal 3 and ground.

[9] i.e., between terminal 3 and ground.

1.3.1 The Closed-Loop Gain

We now wish to analyze the circuit in **Fig.** 1.7 to determine the closed-loop gain (G), which we defined here as:

$$G \equiv \frac{v_0}{v_0}$$

We will do so assuming the op amp to be ideal. Figure 2.6(a) shows the equivalent circuit, and the analysis proceeds as follows:

The gain of the op-amp (A) is very large.[10] If we assume the circuit is operating and producing a finite output voltage at terminal 3, then the voltage between the op-amp input terminals should be negligibly small and ideally zero. Specifically, if we call the output voltage v_0, then, by definition,

[10] ideally infinite.

$$v_2 - v_1 = \frac{v_0}{A} = 0 \tag{1.5}$$

Through this equivalence, it follows that the voltage at the inverting input terminal (v_1) is given by $v_1 = v_2$. That is, because the gain (A) approaches infinity, the voltage v_1 approaches and ideally equals v_2.

> We speak of this as the two (2) input terminals **tracking each other in potential**.

We also speak of a virtual short circuit that exists between the two (2) input terminals. Here the word virtual should be emphasized.[11] A virtual short circuit means that whatever voltage is at 2 will automatically appear at 1 because of the infinite gain (A). However, terminal 2 happens to be connected to ground. Therefore $v_2 = 0$ and $v_1 = 0$. We speak of terminal 1 as being a **virtual ground**.[12] Now that we have determined v_1 we are in a position to apply Ohm's law and find the current i_1 through R_1 (see Fig. 2.6) as follows:

$$i_1 = \frac{v_I - v_1}{R_1} = \frac{v_I - 0}{R_1} = \frac{v_I}{R_1}$$

Where will this current go?

It cannot go into the op amp, since the ideal op amp has an infinite input impedance and hence draws zero current. It follows that i_1 will have to flow through R_2, to the low-impedance terminal 3. We can then apply Ohm's law to R_2 and determine v_0.

That is,

$$\begin{aligned} v_o &= v_1 - i_1 R_2 \\ &= 0 - \frac{v_I}{R_1} R_2 \end{aligned} \qquad \text{therefore} \qquad \frac{v_o}{v_i} = -\frac{R_2}{R_1}$$

which is the required closed-loop gain. Figure 2.6(b) illustrates these steps and indicates by the circled numbers the order in which the analysis is performed. We therefore see the closed-loop gain is simply the ratio of the two (2) resistances R_2 and R_1. The minus sign means that the closed-loop amplifier provides signal inversion. Thus if $R_2/R_1 = 10$ and we apply at the input (v_1) a sine-wave signal of 1 V peak-to-peak, then the output vo will be a sine wave of 10 V peak-to-peak and phase-shifted 180ř with respect to the input sine wave. Because of the minus sign associated with the closed-loop gain, this configuration is called the inverting configuration.

The fact that the closed-loop gain depends entirely on external passive components[13] is very significant. It means that we can make the closed-loop gain as accurate as we want by selecting passive components of appropriate accuracy. It also means that the closed-loop gain is (ideally) independent of the op-amp gain. This is a dramatic illustration of negative feedback:

> We started out with an amplifier having very large gain A, and through applying negative feedback we have obtained a closed-loop gain R_2/R_1 that is much smaller than A but is stable and predictable.

[11] One should **NOT** make the mistake of physically shorting terminals 1 and 2 together while analyzing a circuit.

[12] that is, having zero voltage but not physically connected to ground

[13] Here, these are resistors R_1 and $_2$

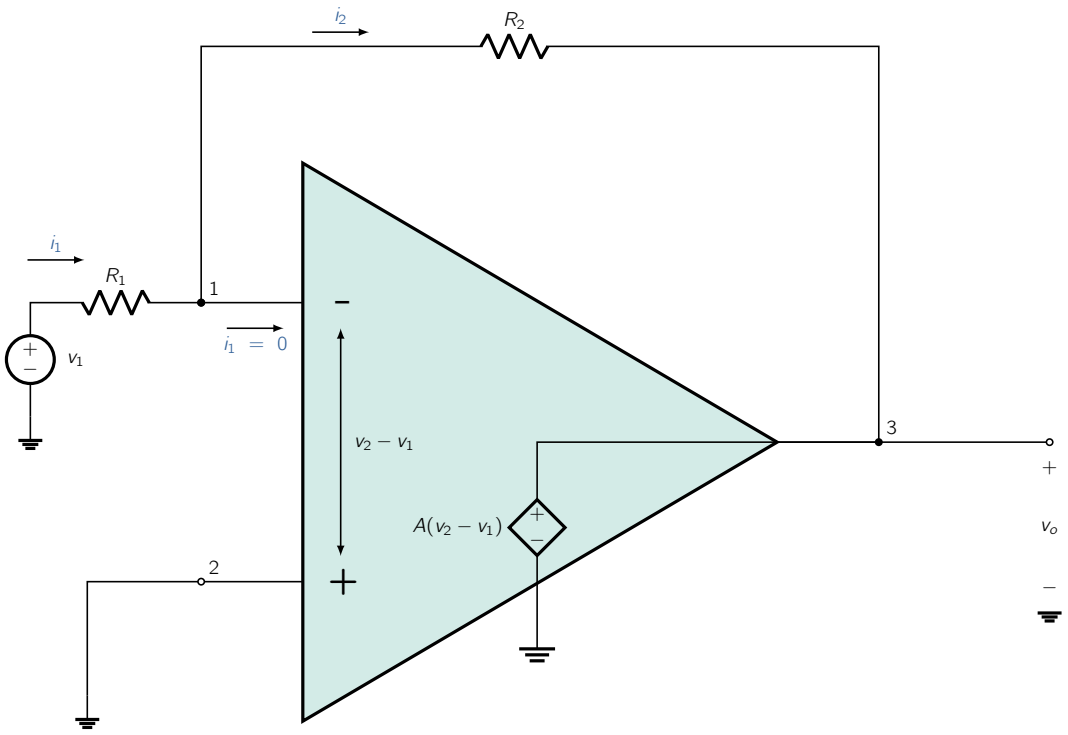


Figure 1.8 Analysis of the inverting configuration taking into account the finite open-loop gain of the op amp.

That is, we are trading gain for accuracy.

1.3.2 Effect of Finite Open-Loop Gain

The points just made are more clearly illustrated by deriving an expression for the closed-loop gain under the assumption that the op-amp open-loop gain A is finite. If we denote the output voltage vo, then the voltage between the two input terminals of the op amp will be v_0/A. Since the positive input terminal is grounded, the voltage at the negative input terminal must be $-v_0/A$. The current i_1 through R_1 can now be found from

$$i_1 = \frac{v_I - (-v_O/A)}{R_1} = \frac{v_I + v_O/A}{R_1}$$

The infinite input impedance of the op amp forces the current i_1 to flow entirely through R_2. The output voltage v_0 can thus be determined from

$$v_0 = -\frac{v_0}{4} - i_1 R_2 = -\frac{v_0}{A} - \left(\frac{v_1 + v_0/A}{R_1}\right) R_2$$

Collecting terms, the closed-loop gain G is found as:

$$G \equiv \frac{v_0}{v_t} = \frac{-R_2/R_1}{1 + (1 + R_2/R_1)/A} \tag{1.6}$$

We note that as A approaches ∞, G approaches the ideal value of $-R_2/R_1$. Also, from **Fig. ??** we see that as A approaches ∞, the voltage at the inverting input terminal approaches zero. This is the virtual-ground assumption we used in our earlier analysis when the op amp was assumed to be ideal. Finally, note that Eq. (**1.6**) in fact indicates that to minimize the dependence of the closed-loop gain G on the value of the open-loop gain A, we should make:

$$1 + \frac{R_2}{R} \ll A$$

1.3.3 Input and Output Resistances

Assuming an ideal op amp with infinite open-loop gain, the input resistance of the closed-loop inverting amplifier of Fig. 2.5 is simply equal to R_1. This can be seen from Fig. 2.6(b), where

$$R_i \equiv \frac{v_I}{i_1} = \frac{v_I}{v_I/R_1} = R_1$$

Now recall that in Section 1.5 we learned that the amplifier input resistance forms a voltage divider with the resistance of the source that feeds the amplifier. Thus, to avoid the loss of signal strength, voltage amplifiers are required to have high input resistance. In the case of the invert- ing op-amp configuration we are studying, to make R_i high we should select a high value for R_1. However, if the required gain R_2/R_1 is also high, then R_2 could become impractically large.[14] We may conclude that the inverting configuration suffers from a low input resistance.

[14] This is in the range of few MΩ.

Since the output of the inverting configuration is taken at the terminals of the ideal voltage source $A(v_2 - v_1)$ (see Fig. 2.6a), it follows that the output resistance of the closed-loop amplifier is zero.

Information: Application: The Weighted Summer

An important application of the **inverting configuration** is the weighted-summer circuit shown in Fig. 2.10. Here we have a resistance R_f, in the negative-feedback path, similar to before. But we have a number of input signals ($v_1, \cdots, v_n$) each applied to a corresponding resistor ($R_1, \cdots, R_n$), which are connected to the **inverting terminal** of the op-amp.

From our previous discussion, the ideal op amp will have a virtual ground appearing at its negative input terminal.

Ohm's law then tells us that the currents ($i_1, \cdots, i_n$) are given by:

$$i_1 = \frac{v_1}{R_1},\ i_2 = \frac{v_2}{R_2},\ \ldots,\ i_n = \frac{v_n}{R_n}$$

All these currents sum together to produce the current i,

$$i = i_1 + i_2 + \cdots + i_n \tag{1.7}$$

which will be forced to flow through R_f.[15] The output voltage (v_o) may now be determined by another application of Ohm's law,

$$v_0 = 0 - iR_f = -iR_f$$

[15] since no current flows into the input terminals of an ideal op amp.

and therefore:

$$v_o = -\left(\frac{R_f}{R_1}v_1 + \frac{R_f}{R_2}v_2 + \cdots + \frac{R_f}{R_n}v_n\right) \tag{1.8}$$

That is, the output voltage is a weighted sum of the input signals $v_1, \cdots, v_n$. This circuit is therefore called a **weighted summer**. Note that each summing coefficient may be independently adjusted by adjusting the corresponding "feed-in" resistor (R_1 to R_n). This nice property, which greatly simplifies circuit adjustment, is a direct consequence of the virtual ground that exists at the inverting op-amp terminal. In the weighted summer of Fig. 2.10 all the summing coefficients must be of the same sign. The need occasionally arises for summing signals with opposite signs. Such a function can be implemented, however, using two op amps as shown in Fig. 2.11. Assuming ideal op amps, it can be easily shown that the output voltage is given by

$$v_1\left(\frac{R_a}{R_a}\right)\left(\frac{R_c}{R_a}\right) + v_2\left(\frac{R_a}{R_a}\right)\left(\frac{R_c}{R_a}\right) - v_3\left(\frac{R_c}{R_a}\right) - v_4\left(\frac{R_c}{R_a}\right) \tag{1.9}$$

Weighted summers are utilized in a variety of applications including in the design of audio systems, where they can be used in mixing signals originating from different musical instruments.

1.4 Non-Inverting Amplifier

The second closed-loop configuration we shall study is shown in **Fig.** 1.9. Here the input signal v_i is applied directly to the positive input terminal of the op-amp while one terminal of R_1 is connected to **ground**.

1.4.1 The Closed-Loop Gain

Analysis of the non-inverting circuit to determine its closed-loop gain (v_0/v_1) is illustrated in **Fig.** 1.9. Again the order of the steps in the analysis is indicated by circled numbers. Assuming the op-amp is ideal with infinite gain, a virtual short circuit exists between its two (2) input terminals.

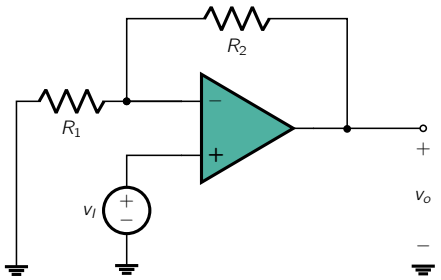


Figure 1.9: The noninverting configuration.

Therefore the difference input signal is:

$$v_{Id} = \frac{v_o}{A} = 0 \qquad \text{for} \qquad A = \infty$$

Therefore the voltage at the inverting input terminal will be equal to that at the non-inverting input terminal, which is the applied voltage v_i. The current through R_1 can then be determined as v_I/R_I. Because of the infinite input impedance of the op amp, this current will flow through R_2, as shown in **Fig. ??**.

Now the output voltage can be determined from

$$v_O = v_I + \left(\frac{v_I}{R_1}\right) R_2 \qquad \text{which gives} \qquad \frac{v_O}{v_I} = 1 + \frac{R_I}{R_2} \tag{1.10}$$

Further insight into the operation of the non-inverting configuration can be obtained by considering the following:

> Since the current into the op-amp inverting input is zero, the circuit composed of R_1 and R_2 acts in effect as a voltage divider feeding a fraction of the output voltage back to the inverting input terminal of the op amp;

that is,

$$v_1 = v_0 \left(\frac{R_1}{R_1 + R_2}\right) \tag{1.11}$$

Then the infinite op-amp gain and the resulting virtual short circuit between the two input terminals of the op amp forces this voltage to be equal to that applied at the positive input terminal; thus,

$$v_o \left(\frac{R_1}{R_1 + R_2}\right) = v_I$$

which produces the gain expression given in Eq. (**1.10**).

This is an appropriate point to reflect further on the action of the negative feedback present in the noninverting circuit of Fig. 2.12. Let vI increase. Such a change in vI will cause vId to increase, and vo will correspondingly increase as a result of the high[16] gain of the op amp. However, a fraction of the increase in vo will be fed back to the inverting input terminal of the op amp through the (R_1, R_2) voltage divider. The result of this feedback will be to counteract the increase in v_{Id}, driving v_{Id} back to zero, albeit at a higher value of v_o that corresponds to the increased value of v_t. This degenerative action of negative feedback gives it the alternative name degenerative feedback. Finally, note that the argument above applies equally well if v_I decreases.

[16] ideally infinite

1.4.2 Effect of Finite Open-Loop Gain

As we have done for the inverting configuration, we now consider the effect of the finite op-amp open-loop gain A on the gain of the noninverting configuration. Assuming the op amp to be ideal except for having a finite open-loop gain A, it can be shown that the closed-loop gain of the noninverting amplifier circuit of Fig. 2.12 is given by

$$G \equiv \frac{v_O}{v_I} = \frac{1 + (R_2/R_1)}{1 + \frac{1+(R_2/R_1)}{4}} \tag{1.12}$$

Observe that the denominator is identical to that for the case of the inverting configuration (Eq. 2.5). This is no coincidence; it is a result of the fact that both the inverting and the noninverting configurations have the same feedback loop, which can be readily seen if the input signal source is eliminated.[17] The numerators, however, are different, for the numerator gives the ideal or nominal closed-loop gain $(-R_2/R_1)$ for the inverting configuration, and $1 + R_2/R_1$ for the noninverting configuration). Finally, we note (with reassurance) that the gain expression in Eq. (2.11) reduces to the ideal value for $A = \infty$.

[17] i.e., short-circuited.

In fact, it approximates the ideal value for

$$A \gg 1 + \frac{R_2}{R_1}$$

This is the same condition as in the inverting configuration, except that here the quantity on the right-hand side is the nominal closed-loop gain. The expressions for the actual and ideal values of the closed-loop gain G in Eqs. (2.11) and (2.9), respectively, can be used to determine the percentage error in G resulting from the finite op-amp gain A as

$$\text{Percent gain error} = -\frac{1 + (R_2/R_1)}{A + 1 + (R_2/R_1)} \times 100 \tag{1.13}$$

Thus, as an example, if an op amp with an open-loop gain of 1000 is used to design a non-inverting amplifier with a nominal closed-loop gain of 10, we would expect the closed-loop gain to be about 1% below the nominal value.

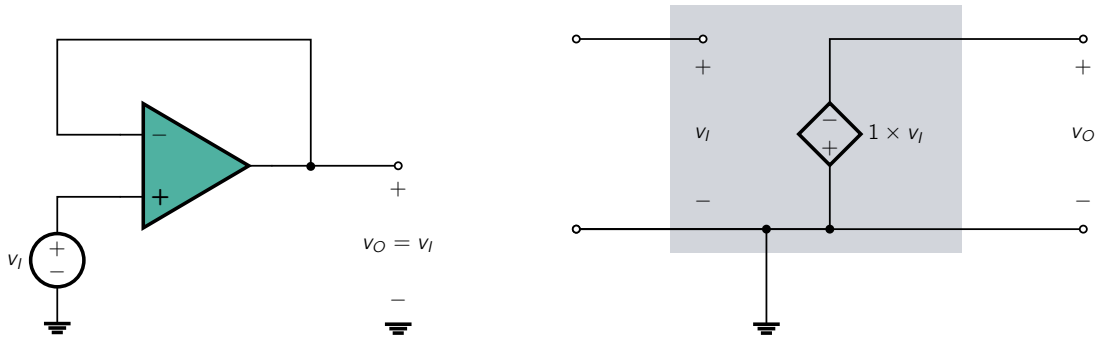


Figure 1.10: (a) The unity-gain buffer or follower amplifier. (b) Its equivalent circuit model.

1.4.3 Input and Output Resistance

The gain of the non-inverting configuration is positive and hence the name noninverting. The input impedance of this closed-loop amplifier is ideally infinite, since no current flows into the positive input terminal of the op amp. The output of the noninverting amplifier is taken at the terminals of the ideal voltage source $A(v_2 - v_1)$,[18] and thus the output resistance of the noninverting configuration is zero.

[18] see the op-amp equivalent circuit in Fig. 2.3

1.4.4 The Voltage Follower

The property of high input impedance is a very desirable feature of the non-inverting configuration. It enables using this circuit as a buffer amplifier to connect a source with a high impedance to a low-impedance load. We discussed the need for buffer amplifiers in Section 1.5. In many applications the buffer amplifier is not required to provide any voltage gain; rather, it is used mainly as an impedance transformer or a power amplifier. In such cases we may make $R_2 = 0$ and $R_1 = \infty$ to obtain the **unity-gain amplifier** shown in Fig. 2.14(a). This circuit is commonly referred to as a **voltage follower**, since the output "follows" the input. In the ideal case, $v_0 = v_I$, $R_{in} = \infty$, $R_{out} = 0$, and the follower has the equivalent circuit shown in Fig. 2.14(b).

Since in the voltage follower circuit the entire output is fed back to the inverting input, the circuit is said to have 100% negative feedback. The infinite gain of the op amp then acts to make $v_{Id} = 0$ and hence $v_0 = v_I$. Since the noninverting configuration has a gain greater than or equal to unity, depending on the choice of R_2/R_1, some prefer to call it "a follower with gain."

1.5 Difference Amplifiers

Having studied the two basic configurations of op-amp circuits together with some of their direct applications, we are now ready to consider a somewhat more involved but very important application. Specifically, we shall study the use of op amps to design difference.

A difference amplifier is one that responds to the difference between the two signals applied at its input and ideally rejects signals that are common to the two inputs. The representation of signals in terms of their differential and common-mode components was given in **Fig.** 1.6. Although ideally the difference amplifier will amplify only the differential input signal $v_{I,d}$ and reject completely the common-mode input signal $v_{I,cm}$, practical circuits will have an output voltage v_O given by

$$v_O = A_d v_{I,d} + A_{cm} v_{I,cm} \tag{1.14}$$

where A_d denotes the amplifier differential gain and A_{cm} denotes its common-mode gain (ideally zero). The efficacy of a differential amplifier is measured by the degree of its rejection of common-mode signals in preference to differential signals. This is usually quantified by a measure known as the **common-mode rejection ratio** (**CMRR**), defined as

$$\text{CMRR} = 20 \log \frac{|A_d|}{|A_{cm}|} \tag{1.15}$$

The need for difference amplifiers arises frequently in the design of electronic systems, especially those employed in instrumentation. As a common example, consider a transducer providing a small (e.g., 1 mV) signal between its two output terminals while each of the two wires leading from the transducer terminals to the measuring instrument may have a large interference signal (e.g., 1 V) relative to the circuit ground. The instrument front end obviously needs a difference amplifier.

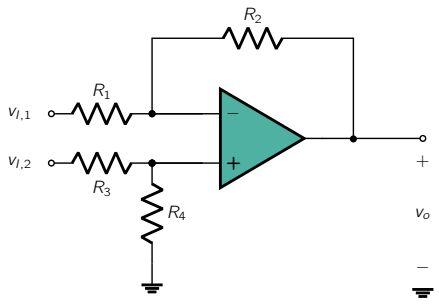
Information: The need for a Difference Configuration

Before we proceed any further we should address a question that the reader might have: The op amp is itself a difference amplifier; why not just use an op amp? The answer is that the very high,ideally infinite, gain of the op amp makes it impossible to use by itself. Rather, as we did before, we have to devise an appropriate feedback network to connect to the op amp to create a circuit whose closed-loop gain is finite, predictable, and stable.

1.5.1 A Single-Op-Amp Difference Amplifier

Our first attempt at designing a difference amplifier is motivated by the observation that the gain of the non-inverting amplifier configuration is positive, $\left(1 + R_2/R_1\right)$, while that of the inverting configuration is negative, $(-R_2/R_1)$. Combining the two configurations together is then a step in the right direction-namely, getting the difference between two input signals.

Figure 1.11
A difference amplifier.



Of course, we have to make the two gain magnitudes equal in order to reject common-mode signals. This, however, can be easily achieved by attenuating the positive input signal to reduce the gain of the positive path from $(1+R_2/R_1)$ to (R_2/R_1). The resulting circuit would then look like that shown in **Fig.** 1.11, where the attenuation in the positive input path is achieved by the voltage divider (R_3, R_4). The proper ratio of this voltage divider can be determined from

$$\frac{R_4}{R_4+R_3}\left(1+\frac{R_2}{R_1}\right)=\frac{R_2}{R_1}$$

which can be put in the form

$$\frac{R_4}{R_4+R_3}=\frac{R_2}{R_2+R_1} \qquad \text{satisfied by} \qquad \frac{R_4}{R_2}=\frac{R_2}{R_1} \tag{1.16}$$

This completes our work. However, Let's step back and verify that the circuit in **Fig.** 1.11 with R_3 and R_4 selected according to Eq. (**1.16**) does in fact function as a difference amplifier. Specifically, we wish to determine the output voltage v_O in terms of v_{I1} and v_{I2}. Toward that end, we observe that the circuit is linear, and thus we can use superposition.

To apply superposition, we first reduce $v_{I,2}$ to zero-that is, ground the terminal to which v_{12} is applied-and then find the corresponding output voltage, which will be due entirely to v_{11}. We denote this output voltage v_{01}. Its value may be found from the circuit in Fig. 2.17(a), which we recognize as that of the inverting configuration. The existence of R_3 and R_4 does not affect the gain expression, since no current flows through either of them. Thus,

$$v_{01}=-\frac{R_2}{R_1}v_{I1} \tag{1.17}$$

Next, we reduce v_{I1} to zero and evaluate the corresponding output voltage v_{02}. The circuit will now take the form shown in Fig. 2.17(b), which we recognize as the noninverting configuration with an

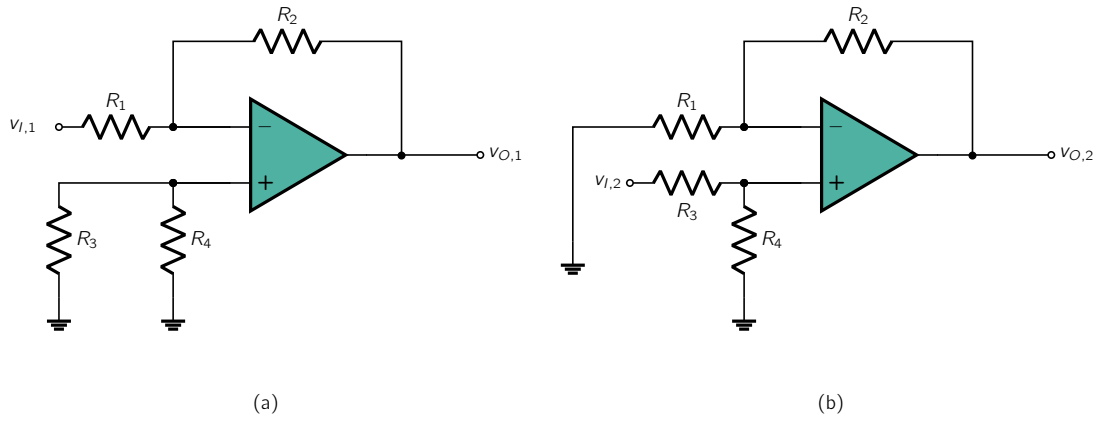


Figure 1.12: Application of superposition to the analysis of the circuit

additional voltage divider, made up of R_3 and R_4, connected to the input v_{12}. The output voltage v_{02} is therefore given by

$$v_{02} = v_{12}\frac{R_4}{R_2+R_4}\left(1+\frac{R_2}{R_1}\right) = \frac{R_2}{R_1}v_{12} \tag{1.18}$$

where we have utilized Eq. (**1.16**). The superposition principle tells us that the output voltage v_0 is equal to the sum of v_{01} and v_{02}. Thus we have

$$v_O = \frac{R_2}{R}(v_{I2}-v_{I1}) = \frac{R_2}{R}v_{Id} \tag{1.19}$$

Thus, as expected, the circuit acts as a difference amplifier with a differential gain A_d of

$$A_d = \frac{R_2}{R_1} \tag{1.20}$$

Of course this is predicated on the op amp being ideal and furthermore on the selection of R_3 and R_4 so that their ratio matches that of R_1 and R_2 (Eq. 2.15). To make this matching requirement a little easier to satisfy, we usually select

$$R_3 = R_1 \qquad \text{and} \qquad R_4 = R_2$$

Let's next consider the circuit with only a common-mode signal applied at the input, as shown in Fig. 2.18. The figure also shows some of the analysis steps. Thus,

$$i_1 = \frac{1}{R_1}\left[v_{Icm} - \frac{R_4}{R_4+R_3}v_{Icm}\right] \tag{1.21}$$

$$= v_{Icm}\frac{R_3}{R_4+R_2}\frac{1}{R_1} \tag{1.22}$$

The output voltage can now be found from

$$v_O = \frac{R_4}{R_4+R_3}v_{Icm} - i_2R_2$$

Substituting $i_2 = i_1$ and for i_1 from Eq. (2.18),

$$v_O = \frac{R_4}{R_4+R_3}v_{Icm} - \frac{R_2}{R_1}\frac{R_3}{R_4+R_3}v_{Icm}$$

$$= \frac{R_4}{R_4+R_2}\left(1-\frac{R_2}{R_1}\frac{R_3}{R_4}\right)v_{Icm}$$

Thus,

$$A_{cm} \equiv \frac{v_O}{v_{cm}} = \left(\frac{R_4}{R_4+R_3}\right)\left(1-\frac{R_2}{R_1}\frac{R_3}{R_4}\right) \tag{1.23}$$

For the design with the resistor ratios selected according to Eq. (2.15), we obtain

$$A_{cm} = 0$$

as expected. Note, however, that any mismatch in the resistance ratios can make A_{cm} nonzero, and hence CMRR finite.

In addition to rejecting common-mode signals, a difference amplifier is usually required to have a high input resistance. To find the input resistance between the two input terminals (i.e., the resistance seen by v_{Id}), called the **differential input resistance** R_{Id}, consider **Fig.** 1.13. Here we have assumed that the resistors are selected so that

$$R_3 = R_1 \qquad \text{and} \qquad R_4 = R_2$$

Now

$$R_{id} \equiv \frac{v_{Id}}{i_I}$$

Since the two input terminals of the op amp track each other in potential, we may write a loop equation and obtain

$$v_{Id} = R_1 i_I + 0 + R_1 i_I$$

Thus,

$$R_{id} = 2R_1 \tag{1.24}$$

Note that if the amplifier is required to have a large differential gain (R_2/R_1), then R_1 of necessity will be relatively small and the input resistance will be correspondingly low, a drawback of this circuit.

Another drawback of the circuit is that it is not easy to vary the differential gain of the amplifier. Both of these drawbacks are overcome in the instrumentation amplifier discussed next.

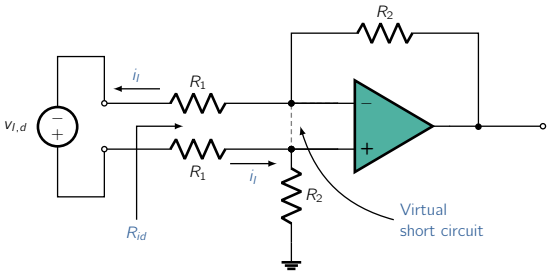


Figure 1.13
Finding the input resis-tance of the difference amplifier for thecase R3= R1 and R4= R2.

1.5.2 A Versatile Circuit: The Instrumentation Amplifier

The low-input-resistance problem of the difference amplifier of Fig. 2.16 can be solved by using voltage followers to buffer the two input terminals; that is, a voltage follower of the type in Fig. 2.14 is connected between each input terminal and the corresponding input terminal of the difference amplifier. However, if we are going to use two additional op amps, we should ask the question: Can we get more from them than just impedance buffering? An obvious answer would be that we should try to get some voltage gain. It is especially interesting that we can achieve this without compromising the high input resistance simply by using followers with gain rather than unity-gain followers. Achieving some or indeed the bulk of the required gain in this new first stage of the differential amplifier eases the burden on the difference amplifier in the second stage, leaving it to its main task of implementing the differencing function and thus rejecting common-mode signals.

The resulting circuit is shown in Fig. 2.20(a). It consists of two stages in cascade. The first stage is formed by op amps A_1 and A_2 and their associated resistors, and the second stage is the by-now-familiar difference amplifier formed by op amp A_3 and its four associated resistors. Observe that as we set out to do, each of A_1 and A_2 is connected in the noninverting configu- ration and thus

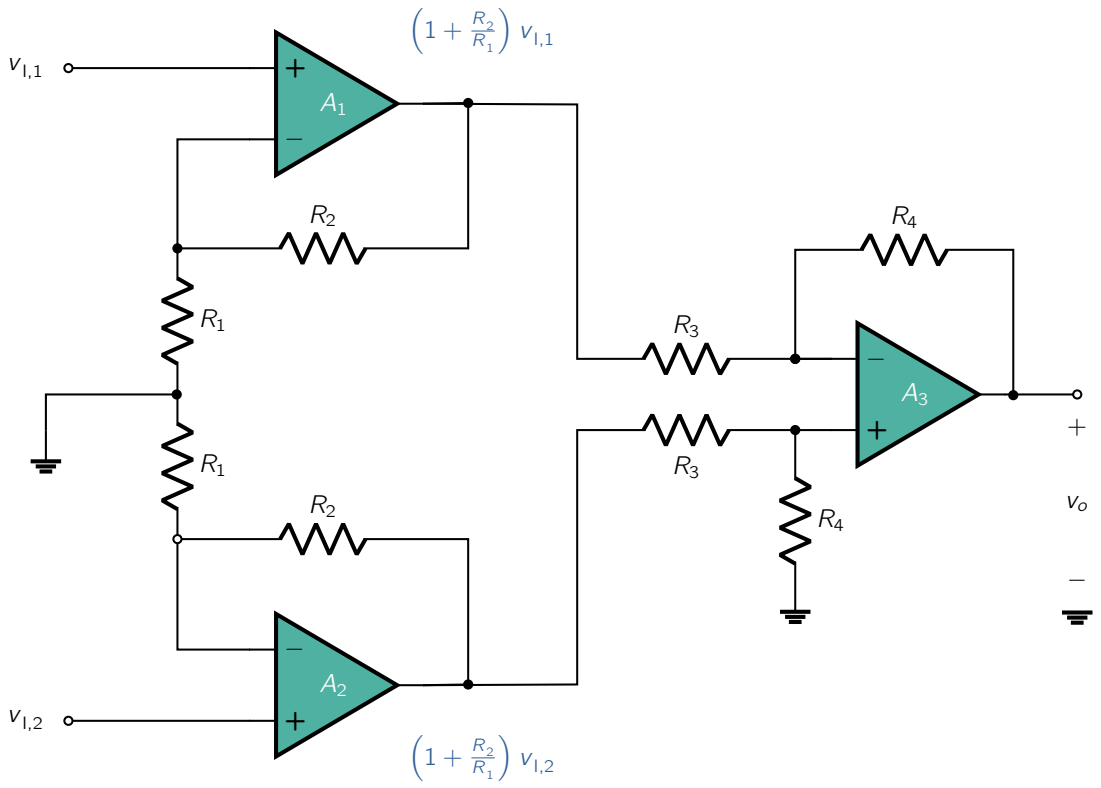


Figure 1.14
A popular circuit for an instrumentation amplifier. (a) Initial approach to the circuit. (b) The circuit in (a) with the connection between node X and ground removed and the two resistors R1 and R1 lumped together. This simple wiring change dramatically improves performance.

realizes a gain of $(1 + R_2/R_1)$. It follows that each of v_{I1} and v_{I2} is amplified by this factor, and the resulting amplified signals appear at the outputs of A_1 and A_2, respectively. The difference amplifier in the second stage operates on the difference signal $(1 + R_2/R_1)(v_{12} - v_{11}) = (1 + R_2/R_1)v_{1d}$ and provides at its output

$$v_O = \frac{R_4}{R_3}\left(1 + \frac{R_2}{R_1}\right)v_{Id}$$

Thus the differential gain realized is

$$A_d = \left(\frac{R_4}{R_3}\right)\left(1 + \frac{R_2}{R_1}\right) \tag{1.25}$$

The common-mode gain will be zero because of the differencing action of the second-stage amplifier

The circuit in Fig. 2.20(a) has the advantage of very high (ideally infinite) input resistance and high differential gain. Also, provided A_1 and A_2 and their corresponding resistors are matched, the two signal paths are symmetric-a definite advantage in the design of a differential amplifier. The circuit, however, has three major disadvantages:

1. The input common-mode signal v_{Im} is amplified in the first stage by a gain equal to that experienced by the differential signal v_{Id}. This is a very serious issue, for it could result in the signals at the outputs of A_1 and A_3 being of such large magnitudes that the op amps saturate (more on op-amp saturation in Section 2.8). But even if the op amps do not saturate, the difference amplifier of the second stage will now have to deal with much larger common-mode signals, with the result that the CMRR of the overall amplifier will inevitably be reduced. 2. The two amplifier channels in

the first stage have to be perfectly matched, otherwise a spurious signal may appear between their two outputs. Such a signal would get amplified by the difference amplifier in the second stage. 3. To vary the differential gain A_d, two resistors have to be varied simultaneously, say the two resistors labeled R_1. At each gain setting the two resistors have to be perfectly matched: a difficult task.

All three problems can be solved with a very simple wiring change: Simply disconnect the node between the two resistors labeled R_1, node X, from ground. The circuit with this small but functionally profound change is redrawn in Fig. 2.20(b), where we have lumped the two resistors (R_1 and R_1) together into a single resistor ($2R_1$).

Analysis of the circuit in Fig. 2.20(b), assuming ideal op amps, is straightforward, as is illustrated in Fig. 2.20(c). The key point is that the virtual short circuits at the inputs of op amps A_1 and A_2 cause the input voltages v_{I1} and v_{I2} to appear at the two terminals of resistor ($2R_1$). Thus the differential input voltage $v_{12} - v_{11} \equiv v_{1d}$ appears across $2R_1$ and causes a current $i = v_{Id}/2R_1$ to flow through $2R_1$ and the two resistors labeled R_2. This current in turn produces a voltage difference between the output terminals of A_1 and A_2 given by

$$v_{O2} - v_{O1} = \left(1 + \frac{2R_2}{2R_1}\right) v_{Id}$$

The difference amplifier formed by op amp A_3 and its associated resistors senses the voltage difference ($v_{02} - v_{01}$) and provides a proportional output voltage v_0:

$$\begin{aligned} v_0 &= \frac{R_4}{R_1}(v_{02} - v_{01}) \\ &= \frac{R_4}{R_2}\left(1 + \frac{R_2}{R_1}\right) v_{Id} \end{aligned}$$

Thus the overall differential voltage-gain is given by

$$A_d \equiv \frac{v_O}{v_H} = \frac{R_4}{R_2}\left(1 + \frac{R_2}{R_1}\right) \tag{1.26}$$

Observe that proper differential operation does not depend on the matching of the two resistors labeled R_2. Indeed, if one of the two is of different value, say R_2', the expression for A_d becomes

$$A_d = \frac{R_4}{R_2}\left(1 + \frac{R_2 + R_2'}{2R_1}\right) \tag{1.27}$$

Consider next what happens when the two input terminals are connected together to a common-mode input voltage v_{Im}. It is easy to see that an equal voltage appears at the negative input terminals of A_1 and A_2, causing the current through $2R_1$ to be zero. Thus there will be no current flowing in the R_2 resistors, and the voltages at the output terminals of A_1 and A_2 will be equal to the input (i.e., v_{Icm}). Thus the first stage no longer amplifies v_{Icm}; it simply propagates v_{Icm} to its two output terminals, where they are subtracted to produce a zero common-mode output by A_3. The difference amplifier in the second stage, however, now has a much improved situation at its input: The difference signal has been amplified by $(1 + R_2/R_1)$ while the common-mode voltage remained unchanged.

Finally, we observe from the expression in Eq. (2.22) that the gain can be varied by changing only one resistor, $2R_1$. We conclude that this is an excellent differential amplifier circuit and is widely

employed as an instrumentation amplifier, that is, as the input amplifier used in a variety of electronic instruments.

1.6 Integrators and Differentiators

The applications we have studied up to now have **only** utilised resistors in the op-amp feedback path and in connecting the signal source to the circuit, that is, in the feed-in path. As a result, circuit operation has been[19] independent of frequency. By allowing the use of both capacitors along with resistors in the feedback and feed-in paths of op-amp circuits, we can extend the use of op-amps and create wide range of useful and exciting applications.

[19] ideally

We begin our study of op-amp RC circuits by considering two (2) basic applications:

signal **integrators** and **differentiators**.

1.6.1 The Inverting Configuration with General Impedances

To start, let us consider the inverting closed-loop configuration with impedances as $Z_1(s)$ and $Z_2(s)$ replacing the previously used resistors R_1 and R_2, respectively.

The resulting circuit is shown in **Fig.** 1.15 and, for an ideal op amp, has the closed-loop gain or, more appropriately, the closed-loop transfer function

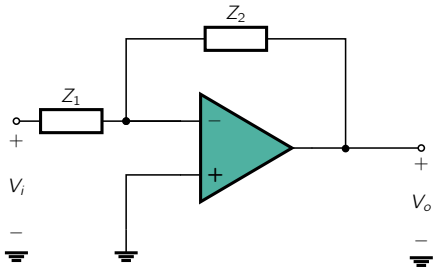


Figure 1.15 **The inverting configuration with general impedances in the feedback and the feed-in paths.**

$$\frac{V_o(s)}{V_i(s)} = -\frac{Z_2(s)}{Z_1(s)} \qquad (1.28)$$

As explained in the previous chapter, replacing s by $\mathbf{j}\omega$ provides the transfer function for physical frequencies ω:

the transmission magnitude and phase for a sinusoidal input signal of frequency (ω).

1.6.2 The Inverting Integrator

By placing a capacitor in the feedback path[20] and a resistor at the input, in place of Z_1, we obtain the circuit as seen in Fig. 2.24(a). We will now see that this circuit realizes the mathematical operation of integration.

[20] i.e., in place of Z_2 in **Fig.** 1.15

Let the input be a time-varying function of $v_1(t)$. The virtual ground at the inverting op-amp input causes $v_1(t)$ to appear in effect across R, and therefore the current $i_1(t)$ will be $v_1(t)/R$. This current flows through the capacitor C, causing charge to accumulate on C. If we assume that the circuit begins operation at time $t = 0$, then at an arbitrary time t, the current $i_1(t)$ will have deposited on C a charge equal to $\int_0^t i_1(t)dt$.

Therefore the capacitor voltage $v_c(t)$ will change by:

$$\frac{1}{C}\int_0^t i_1(t)\ \mathrm{d}t.$$

If the initial voltage on C (at $t = 0$) is denoted V_C, then

$$v_c(t) = V_c + \frac{1}{C}\int_0^t i_1(t)dt$$

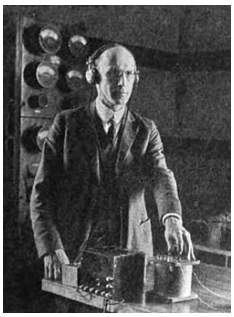
Now the output voltage $v_o(t) = -v_C(t)$; thus,

$$v_O(t) = -\frac{1}{CR}\int_0^t v_I(t)dt - V_C \qquad (1.29)$$

Therefore the circuit provides an output voltage which is proportional to the time integral of the input, with V_C being the initial condition of integration and CR the integrator time constant.

As expected, there is a negative sign attached to the output voltage, and thus this integrator circuit is said to be an **inverting integrator**.

It is also known as a Miller integrator[21] after an early worker in this field.



[21] Noted American electrical engineer, best known for discovering the Miller effect and inventing fundamental circuits for quartz crystal oscillators.

The operation of the integrator circuit can be described alternatively in the frequency domain by substituting $Z_1(s) = R$ and $Z_2(s) = 1/sC$ in Eq. (**1.28**) to obtain the transfer function:

$$\frac{V_o(s)}{V(s)} = -\frac{1}{sCR} \qquad (1.30)$$

For physical frequencies, $s = \mathbf{j}\omega$ and

$$\frac{V_o(\mathbf{j}\omega)}{V_i(\mathbf{j}\omega)} = -\frac{1}{\mathbf{j}\omega CR} \qquad (1.31)$$

Thus the integrator transfer function has magnitude

$$\left|\frac{V_o}{V}\right| = \frac{1}{\omega CR} \qquad \text{and phase} \qquad \phi = +90^\circ \qquad (1.32)$$

The Bode plot for the integrator magnitude response can be obtained by noting from Eq. (**1.32**) that as ω doubles[22] the magnitude is halved (decreased by 6 dB). Therefore the Bode plot is a straight line of slope -6 dB/octave (or, equivalently, -20 dB/ decade). This line (shown in **Fig.** 1.16) intercepts the 0-dB line at the frequency that makes $|V_a/V_i| = 1$, which from Eq. (**1.32**) is:

[22] increases by an octave.

$$\omega_{int} = \frac{1}{CR} \qquad (1.33)$$

The frequency ω_{int} is known as the **integrator frequency** and is simply the inverse of the integrator time constant.

Comparison of the frequency response of the integrator to that of an STC low-pass network indicates that the integrator behaves as a low-pass filter with a corner frequency of zero. Observe also that

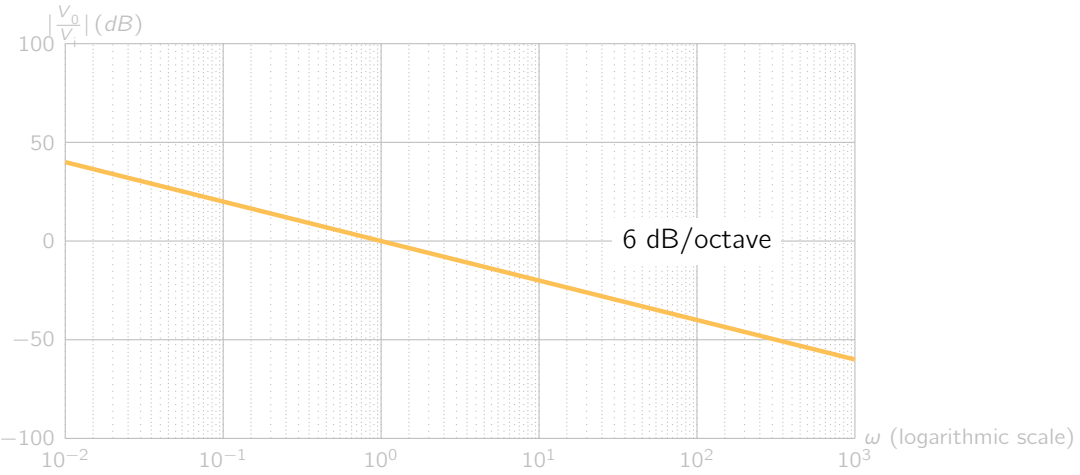


Figure 1.16: Frequency response of the integrator.

at $\omega = 0$, the magnitude of the integrator transfer function is infinite. This indicates that at dc the op amp is operating with an open loop. This should also be obvious from the integrator circuit itself. Reference to Fig. 2.24(a) shows that the feedback element is a capacitor, and thus at dc, where the capacitor behaves as an open circuit, there is no negative feedback! This is a very significant observation and one that indicates a source of problems with the integrator circuit: Any tiny dc component in the input signal will theoretically produce an infinite output. Of course, no infinite output voltage results in practice; rather, the output of the amplifier saturates at a voltage close to the op-amp positive or negative power supply (L_+ or L_-), depending on the polarity of the input dc signal.

The dc problem of the integrator circuit can be alleviated by connecting a resistor R_F across the integrator capacitor C, as shown in **Fig.** 1.17, and therefore the gain at dc will be $-R_F/R$ rather than infinite. Such a resistor provides a DC feedback path. Unfortunately, however, the integration is no longer ideal, and the lower the value of R_F, the less ideal the integrator circuit becomes.

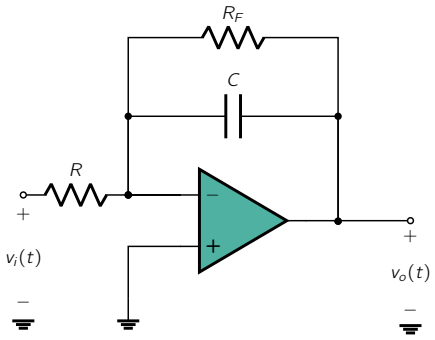


Figure 1.17
The Miller integrator with a large resistance RF connected in parallel with C in order to provide negative feedback and hence finite gain at dc.

> This is because R_F causes the frequency of the integrator pole to move from its ideal location at $\omega = 0$ to one determined by the corner frequency of the STC network (R_F, C).

Specifically, the integrator transfer function becomes

$$\frac{V_o(s)}{V_i(s)} = -\frac{R_F/R}{1 + sCR_F} \qquad (1.34)$$

as opposed to the ideal function of $-1/sCR$. The lower the value we select for RF, the higher the corner frequency $(1/CRF)$ will be and the more nonideal the integrator becomes. Therefore selecting a value for R_F presents the designer with a trade-off between dc performance and signal performance. The effect of R_F on integrator performance is investigated further in

The preceding example hints at an important application of integrators, namely, their use in providing triangular waveforms in response to square-wave inputs.

1.6.3 The Op-Amp Differentiator

Interchanging the location of the capacitor and the resistor of the integrator circuit results in the circuit in Fig. 2.27(a), which performs the mathematical function of differentiation. To see how this comes about, let the input be the time-varying function $vI(t)$, and note that the virtual ground at the inverting input terminal of the op amp causes $vI(t)$ to appear in effect across the capacitor C. Thus the current through C will be $C(dvI/dt)$, and this current flows through the feedback resistor R providing at the op-amp output a voltage $v_o(t)$,

$$v_o(t) = -CR\frac{dv_i(t)}{dt} \tag{1.35}$$

The frequency-domain transfer function of the differentiator circuit can be found by substituting in Eq. (2.24), $Z_1(s) = 1/sC$ and $Z_2(s) = R$ to obtain

$$\frac{V_o(s)}{V_.(s)} = -sCR \tag{1.36}$$

which for physical frequencies $s = j\omega$ yields

$$\frac{V_o(j\omega)}{V_i(j\omega)} = -j\omega CR \tag{1.37}$$

Thus the transfer function has magnitude

$$\left|\frac{V_o}{V}\right| = \omega CR \tag{1.38}$$

and phase

$$\phi = -90^\circ \tag{1.39}$$

The Bode plot of the magnitude response can be found from Eq. (2.34) by noting that for an octave increase in ω, the magnitude doubles (increases by 6 dB). Thus the plot is simply a straight line of slope +6 dB/octave (or, equivalently, +20 dB/decade) intersecting the 0-dB line (where $|V_o/V_i| = 1$) at $\omega = 1/CR$, where CR is the **differentiator time constant** [see Fig. 2.27(b)].

The frequency response of the differentiator can be thought of as the response of an STC high-pass filter with a corner frequency at infinity (refer to Fig. 1.24). Finally, we should note that the very nature of a differentiator circuit causes it to be a "noise magnifier." This is due to the spike introduced at the output every time there is a sharp change in $v1(t)$; such a change could be interference coupled electromagnetically ("picked up") from adjacent signal sources. For this reason

and because they suffer from stability problems (Chapter 11), differentiator circuits are generally avoided in practice. When the circuit of Fig. 2.27(a) is used, it is usually necessary to connect a small-valued resistor in series with the capacitor. This modification, unfortunately, turns the circuit into a nonideal differentiator.

1.7 DC Imperfections

Up to now we have considered the op-amp to be ideal.

> The only exception has been a brief discussion of the effect of the op-amp finite gain (A) both the closed-loop gain of the **inverting** and **non-inverting** configurations.

While for many applications the assumption of an ideal op amp is **NOT** a bad practice on itself, a circuit designer needs to be thoroughly familiar with the characteristics of practical op-amps and the effects of such characteristics on the performance of op-amp circuits.[23]

[23] Only then will the designer be able to use the op amp intelligently, especially if the application at hand is not a straightforward one.

The non-ideal properties of op amps will, of course, limit the operation range of the circuits we have analysed previously. Now, we will consider some of the important non-ideal properties of the op amp. We do this by treating one non-ideal case at a time, beginning with the dc problems to which op-amps are susceptible.

1.7.1 Offset Voltage

As op-amps are direct-coupled devices with large gains at dc, they are prone to dc problems. The first such problem is the dc offset voltage.

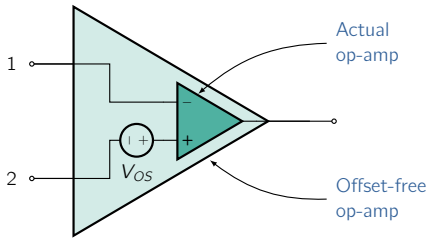


Figure 1.18 **Circuit model for an op amp with input offset voltage V_{OS}.**

To understand this problem consider the following conceptual experiment:

> If the two (2) input terminals of the op amp are tied together and connected to ground (GND), we will see, despite the fact that $v_{Id} = 0$, a there will be a finite dc voltage at the output.

In fact, if the op-amp has a high dc gain, the output will be at either the positive or negative saturation level. The op-amp output can be brought back to its ideal value of 0 V by connecting a DC voltage source of appropriate polarity and magnitude between the two (2) input terminals of the op amp. This external source balances out the input offset voltage of the op amp.

> The input offset voltage (V_{OS}) must be of equal magnitude and of opposite polarity to the voltage we applied externally.

The input offset voltage arises as a result of the unavoidable mismatches present in the input differential stage inside the op-amp. Here, our concern is to investigate the effect of V_{OS} on the operation of closed-loop op-amp circuits.[24]

[24] From a practical perspective, general-purpose op amps exhibit V_{OS} in the range of 1 mV to 5 mV

The value of V_{OS} depends on temperature. The op-amp data sheets generally specify typical and

maximum values for V_{OS} at room temperature as well as the temperature coefficient of V_{OS}, usually in $\mu \cdot V \cdot {}^{\circ}C^{-1}$. They do not, however, specify the polarity of V_{OS} because the component mismatches that give rise to V_{OS} are obviously not known a priori; different units of the same op-amp type may exhibit either a positive or a negative V_{OS}.

To analyze the effect of V_{OS} on the operation of op-amp circuits, we need a circuit model for the op amp with input offset voltage. Such a model is shown in **Fig.** 1.18. It consists of a de source of value V_{OS} placed in series with the positive input lead of an offset-free op amp. The justification for this model follows from the description above.

Analysis of op-amp circuits to determine the effect of the op-amp V_{OS} on their performance is straightforward: The input voltage signal source is short-circuited and the op amp is replaced with the model of **Fig.** 1.18.[25] Following this procedure, we find that both the inverting and the noninverting amplifier configurations result in the same circuit, that shown in Fig. 2.29, from which the output dc voltage due to V_{OS} is found to be

$$V_O = V_{OS}\left[1 + \frac{R_2}{R_1}\right] \tag{1.40}$$

This output dc voltage can have a large magnitude.

> As an example, a non-inverting amplifier with a closed-loop gain of 1000, when constructed from an op-amp with a 5 mV input offset voltage, will have a DC output voltage of +5 V or -5 V[26] rather than the ideal value of 0 V.

[25] Eliminating the input signal, done to simplify matters, is based on the principle of superposition.

[26] The output polarity of course depends on the polarity of V_{OS}.

Now, when an input signal is applied to the amplifier, the corresponding signal output will be superimposed on the 5 V DC. Obviously then, the allowable signal swing at the output will be reduced.

> Even worse, if the signal to be amplified is DC, we would not know whether the output is due to V_{OS} or to the signal.

Some op amps are provided with two (2) additional terminals to which a specified circuit can be connected to trim to zero the output DC voltage due to V_{OS}. **Fig.** 1.19 shows such an arrangement that is typically used with general-purpose op amps. A potentiometer is connected between the offset-nulling terminals with the wiper of the potentiometer connected to the op-amp negative supply. Moving the potentiometer wiper introduces an imbalance that counteracts the asymmetry present in the internal op-amp circuitry and that gives rise to V_{OS}. It should be noted, however, that even though the dc output offset can be trimmed to zero, the problem remains of the variation (or drift) of V_{OS} with temperature.

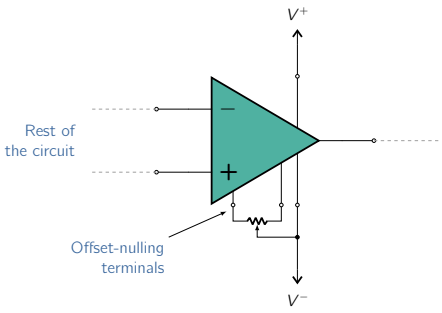


Figure 1.19
The output dc offset voltage of an op amp can be trimmed to zero by connecting a potentiometer to the two offset-nulling terminals. The wiper of the potentiometer is connected to the negative supply of the op amp.

One way to overcome the dc offset problem is by capacitively coupling the amplifier. This, however,

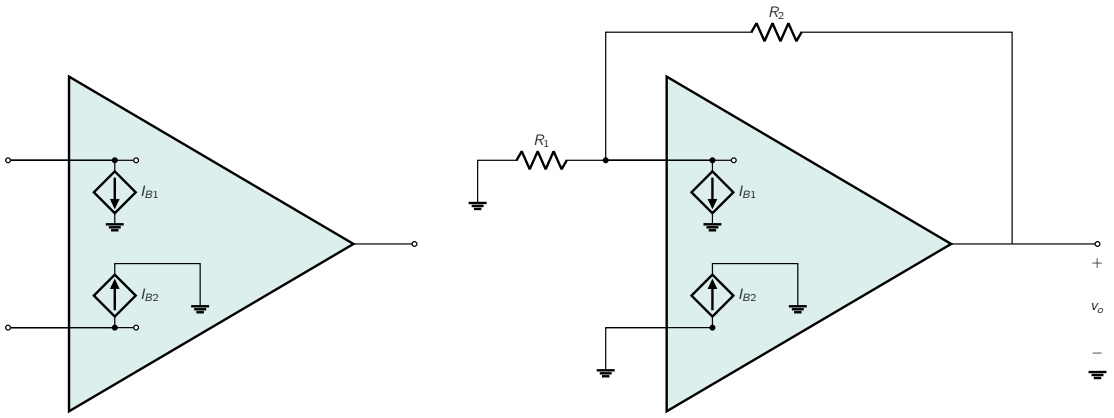


Figure 1.20: (a) The op-amp input bias currents represented by two current sources IB1 and IB2 (b) Analysis of the closed-loop amplifier, taking into account the input bias currents.

will be possible only in applications where the closed-loop amplifier is not required to amplify dc or very-low-frequency signals. Figure 2.31(a) shows a capacitively coupled amplifier. Because of its infinite impedance at dc, the coupling capacitor will cause the gain to be zero at dc. As a result, the equivalent circuit for determining the dc output voltage resulting from the op-amp input offset voltage V_{OS} will be that shown in Fig. 2.31(b).

Therefore V_{OS} sees in effect a unity-gain voltage follower, and the dc output voltage V_0 will be equal to V_{OS} rather than $V_{OS}(1 + R_2/R_1)$, which is the case without the coupling capacitor. As far as input signals are concerned, the coupling capacitor C forms together with R_1 an STC high-pass circuit with a corner frequency of $\omega_0 = 1/CR_1$. Thus the gain of the capacitively coupled amplifier will fall off at the low-frequency end [from a magnitude of $(1 + R_2/R_1)$ at high frequencies] and will be 3 dB down at ω_0.

1.7.2 Input Bias and Offset Currents

The second DC problem encountered in op amps is illustrated in **Fig.** 1.20a. For the op amp to operate, its two input terminals have to be supplied with dc currents, termed the input bias currents.[27] In **Fig.** 1.20a these two currents are represented by two current sources, I_{B1} and I_{B2}, connected to the two input terminals. It should be emphasized that the input bias currents are independent of the fact that a real op amp has finite (though large) input resistance (not shown in **Fig.** 1.20a).

[27] This is the case for op amps constructed using bipolar junction transistors (BJTs). Those using MOSFETs in the first (input) stage do not draw an appreciable input bias current; nevertheless, the input terminals should have continuous dc paths to ground. More on this in later chapters.

> The op-amp manufacturer usually specifies the average value of I_{B1} and I_{B2} as well as their expected difference. The average value I_B is called the **input bias current**,

$$I_B = \frac{I_{B1} + I_{B2}}{2}$$

and the difference is called the input offset current and is given by

$$I_{OS} = |I_{B1} - I_{B2}|$$

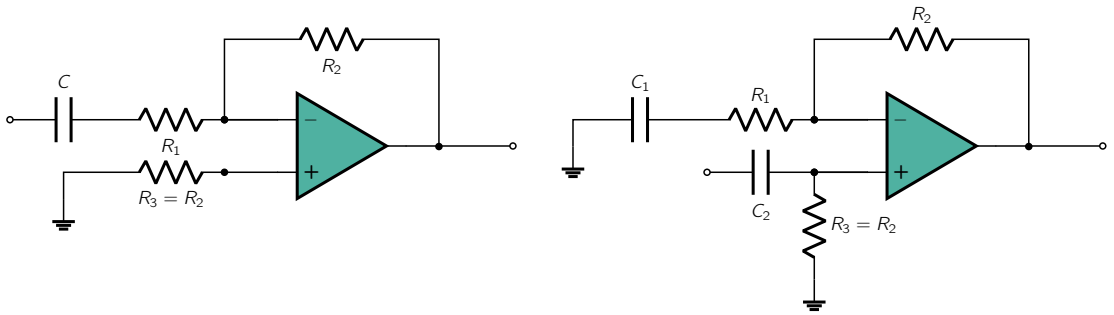


Figure 1.21: (a)In an ac-coupled amplifier the dc resistance seen by the inverting terminal is R2 ; hence R3 is chosen equal to R2. (b) Illustrating the need for a continuous dc path for each of the op-amp input terminals. Specifically, note that the amplifier will not work without resistor R3.

Typical values for general-purpose op amps that use bipolar transistors are $IB = 100$ nA and $I_{OS} = 10$ nA.

We now wish to find the dc output voltage of the closed-loop amplifier due to the input bias currents. To do this we ground the signal source and obtain the circuit shown in **Fig.** 1.20b for both the inverting and noninverting configurations. As shown in **Fig.** 1.20b, the output dc voltage is given by

$$V_0 = I_{B1}R_2 \simeq I_B R_2 \tag{1.41}$$

This obviously places an upper limit on the value of R_2. Fortunately, however, a technique exists for reducing the value of the output dc voltage due to the input bias currents. The method consists of introducing a resistance R_3 in series with the noninverting input lead, From a signal point of view, R_3 has a negligible effect.[28] The appropriate value for R_3 can be determined by analyzing the circuit in Fig. 2.34, where analysis details are shown, and the output voltage is given by

[28] ideally no effect

$$V_\alpha = -I_\alpha R_3 + R_2(I_{\alpha 1} - I_{\alpha 2}R_3/R_1) \tag{1.42}$$

Consider first the case $I_{B1} = I_{B2} = I_B$, which results in

$$V_o = I_R[R_2 - R_3(1 + R_2/R_1)]$$

Thus we can reduce V_0 to zero by selecting R_3 such that

$$R_3 = \frac{R_2}{1 + R_2/R_1} = \frac{R_1 R_2}{R_1 + R_2} \tag{1.43}$$

That is, R_3 should be made equal to the parallel equivalent of R_1 and R_2. Having selected R_3 as above, let us evaluate the effect of a finite offset current I_{OS}. Let $I_{B1} = I_B + I_{OS}/2$ and $I_{B2} = I_B - I_{OS}/2$, and substitute in Eq. (2.38). The result is

$$V_o = I_{os}R_2 \tag{1.44}$$

which is usually about an order of magnitude smaller than the value obtained without R_3 (Eq. (**1.41**)). We can therefore conclude that to minimize the effect of the input bias currents, one should place in the positive lead a resistance equal to the equivalant dc resistance seen by the inverting terminal.

We emphasize the word *dc* in the last statement; note that if the amplifier is ac-coupled, we should select $R_3 = R_2$, as shown in Fig. 2.35.

While we are on the subject of ac-coupled amplifiers, we should note that one must always provide a continuous dc path between each of the input terminals of the op amp and ground. This is the case no matter how small I_B is. For this reason the ac-coupled noninverting amplifier of Fig. 2.36 will not work without the resistance R_3 to ground.

> Unfortunately, including R_3 lowers considerably the input resistance of the closed-loop amplifier.

1.7.3 Effects of Offset of Inverting Integrator

Our discussion of the inverting integrator circuit mentioned the susceptibility of this circuit to saturation in the presence of small dc voltages or currents and therefore we have to consider the effect of the op-amp DC offsets on its operation. As will be seen, these effects can be quite dramatic.

To see the effect of the input dc offset voltage V_{OS}, consider the integrator circuit in Fig. 2.37, where for simplicity we have short-circuited the input signal source. Analysis of the circuit is straightforward and is shown in Fig. 2.37. Assuming for simplicity that at time $t = 0$ the voltage across the capacitor is zero, the output voltage as a function of time is given by

$$v_O = V_{OS} + \frac{V_{OS}}{CR} t \tag{1.45}$$

Therefore v_O increases linearly with time until the op amp saturates. As should be expected, the dc input offset current I_{OS} produces a similar problem. Figure 2.38 illustrates the situation. Observe that we have added a resistance R in the op-amp positive-input lead in order to keep the input bias current IB from flowing through C. Nevertheless, the offset current I_{OS} will flow through C and cause v_O to ramp linearly with time until the op amp saturates.

As mentioned previously on the study of integrator, the dc problem of the integrator circuit can be alleviated by connecting a resistor R_F across the integrator capacitor C, as shown in Fig. 2.25. Such a resistor provides a dc path through which the dc currents (V_{OS}/R) and I_{OS} can flow,[29] with the result that v_0 will now have a dc component $[V_{OS}(1 + R_F/R) + I_{OS}R_F]$ instead of rising linearly. To keep the dc offset at the output small, one would select a low value for RF.

[29] assuming a resistance equal to $R \parallel R_F$ is connected in the positive op-amp lead

> Unfortunately, however, the lower the value of R_F, the less ideal the integrator circuit becomes.

1.8 Finite Open-Loop Gain and Bandwidth

1.8.1 Frequency Dependence of the Open-Loop Gain

The differential open-loop gain (A) of an op amp is **NOT** infinite. It is finite and decreases with frequency. Figure 2.39 shows a plot for $|A|$, with the numbers typical of some commercially available general-purpose op amps.

> although the gain is quite high at dc and low frequencies, it starts to fall off at a rather low frequency (10 Hz in our example).

The uniform -20-dB/decade gain rolloff shown is typical of **internally compensated** op amps. These are units that have a network[30] included within the same IC chip whose function is to cause the op-amp gain to have the single-time-constant (STC) low-pass response shown. This process of modifying the open-loop gain is termed frequency compensation, and its purpose is to ensure that op-amp circuits will be stable (as opposed to oscillatory).

[30] usually a single capacitor

By analogy to the response of low-pass STC circuits, the gain $A(s)$ of an internally compensated op amp may be expressed as

$$A(s) = \frac{A_0}{1 + s/\omega_b} \tag{1.46}$$

which for physical frequencies, $s = \mathbf{j}\omega$, becomes the following:

$$A(\mathbf{j}\omega) = \frac{A_0}{1 + \mathbf{j}\omega/\omega_b} \tag{1.47}$$

where A_0 denotes the dc gain and ω_b is the 3-dB frequency.[31] For the example shown in Fig. 2.39, $A_0 = 10^5$ and $\omega_b = 2\pi \times 10$ rad/s. For frequencies $\omega \gg \omega_b$ (about 10 times and higher) Eq. (2.43) may be approximated by

[31] corner frequency or "break" frequency

$$A(\mathbf{j}\omega) \simeq \frac{A_0\omega_b}{\mathbf{j}\omega} \qquad \text{therefore} \qquad |A(\mathbf{j}\omega)| = \frac{A_0\omega_b}{\omega} \tag{1.48}$$

from which it can be seen that the gain |A| reaches unity (0 dB) at a frequency denoted by ω and given by

$$\omega_t = A_0\omega_b \tag{1.49}$$

Substituting in Eq. (**1.48**) gives:

$$A(\mathbf{j}\omega) \simeq \frac{\omega_t}{i\omega} \tag{1.50}$$

The frequency $f_t = \omega_t/2\pi$ is usually specified on the data sheets of commercially available op amps and is known as the **unity-gain bandwidth**. Also note that for $\omega \gg \omega_b$ the open-loop gain in Eq. (**1.46**) becomes:

$$A(s) \simeq \frac{\omega_t}{s} \tag{1.51}$$

The gain magnitude can be obtained from Eq. (**1.50**) as

$$|A(\mathbf{j}\omega)| \simeq \frac{\omega_t}{\omega} = \frac{f_t}{f} \tag{1.52}$$

Therefore if f_t is known (10⁶ Hz in our example), one can easily determine the magnitude of the op-amp gain at a given frequency f. Furthermore, observe that this relationship correlates with the Bode plot in Fig. 2.39. Specifically, for $f \gg f_b$, doubling f (an octave increase) results in halving the gain (a 6-dB reduction). Similarly, increasing f by a factor of 10 (a decade increase) results in reducing $|A|$ by a factor of 10 (20 dB).

As a matter of practical importance, we note that the production spread in the value of f_t between op-amp units of the same type is usually much smaller than that observed for A_0 and fb. For this reason ft is preferred as a specification parameter. Finally, it should be mentioned that an op amp having this uniform -6-dB/octave (or equivalently -20-dB/decade) gain rolloff is said to have a **single-pole model**. Also, since this single pole dominates the amplifier frequency response, it is called a dominant pole. For more on poles (and zeros), the reader may wish to consult Appendix F.

1.8.2 Frequency Response of Closed-Loop Amplifiers

We next consider the effect of limited op-amp gain and bandwidth on the closed-loop transfer functions of the two (2) basic configurations:

- the inverting circuit of Fig. 2.5, and the
- noninverting circuit of Fig. 2.12.

The closed-loop gain of the inverting amplifier, assuming a finite op-amp open-loop gain A, was derived in Eq. (**1.6**), which we repeat here as:

$$\frac{V_o}{V} = \frac{-R_2/R_1}{1 + (1 + R_2/R_1)/A} \tag{1.53}$$

Substituting for A from Eq. (**1.46**) and using Eq. (**1.49**) gives

$$\frac{V_o(s)}{V_i(s)} = \frac{-R_2/R_1}{1 + \frac{1}{A_o}\left(1 + \frac{R_2}{R}\right) + \frac{s}{\omega l(1+R_2/R_1)}} \tag{1.54}$$

For $A_0 \gg 1 + R_2/R_1$, which is usually the case,

$$\frac{V_o(s)}{V_i(s)} \simeq \frac{-R_2/R_1}{1 + \frac{s}{\omega l(1+R/lR_1)}} \tag{1.55}$$

which is of the same form as that for a low-pass STC network (see Table 1.2, page 36). Thus the inverting amplifier has an STC low-pass response with a dc gain of magnitude equal to R_2/R_1. The closed-loop gain rolls off at a uniform -20-dB/decade slope with a corner frequency (3-dB frequency) given by

$$\omega_{3\text{dB}} = \frac{\omega_t}{1 + R_2/R_1} \tag{1.56}$$

Similarly, analysis of the noninverting amplifier of Fig. 2.12, assuming a finite open-loop gain A, yields the closed-loop transfer function

$$\frac{V_o}{V_i} = \frac{1 + R_2/R_1}{1 + (1 + R_2/R_1)/A} \tag{1.57}$$

Substituting for A from Eq. (2.42) and making the approximation $A_0 \gg 1 + R_2/R_1$ results in

$$\frac{V_o(s)}{V_i(s)} \simeq \frac{1 + R_2/R_1}{1 + \frac{s}{\omega/(1+R_2/R_1)}} \tag{**1.58**}$$

Thus the noninverting amplifier has an STC low-pass response with a dc gain of $(1 + R_2/R_1)$ and a 3-dB frequency given also by Eq. (2.53).

The table in Example 2.6 above clearly illustrates the trade-off between gain and bandwidth: For a given op amp, the lower the closed-loop gain required, the wider the bandwidth achieved. Indeed, the noninverting configuration exhibits a constant **gain-bandwidth product** equal to f_t of the op amp. An interpretation of these results in terms of feedback theory will be given in Chapter 11.

1.9 Large-Signal Operation of Op Amps

In our final section for studying op-amps, we will look at the limitations on the performance of op-amp circuits when large output signals are present.

1.9.1 Output Voltage Saturation

Similar to all other amplifiers, op-amps operate linearly over a limited range of output voltages. Specifically, the op-amp output saturates in the manner shown discussed in the previous chapter with L_+ and L_- within 1 V or so of the positive and negative power supplies, respectively.

Therefore, an op-amp operating from ±15 V supplies will saturate when the output voltage reaches about +13 V in the positive direction and −13 V in the negative direction. For this particular op amp the rated output voltage is said to be ±13 V.

> To avoid clipping off the peaks of the output waveform, and the resulting waveform distortion, the input signal must be kept correspondingly small.

1.9.2 Output Current Limits

Another limitation on the operation of op-amps is their output current is limited to a specified maximum.

> For instance, the popular 741 op amp is specified to have a maximum output current[32] of ±20 mA.

Therefore, in designing closed-loop circuits utilizing the 741, the designer[33] has to ensure that under no condition will the op-amp be required to supply an output current, in either direction, exceeding ±20 mA.

This, of course, has to include both the current in the feedback circuit as well as the current supplied to a load resistor. If the circuit requires a larger current, the op-amp output voltage will saturate at the level corresponding to the maximum allowed output current.

[32] This is the maximum amount of current the op amp can source or sink without damage. Exceeding this limit can damage the device.

[33] That is you.

1.9.3 Slew Rate

Another phenomenon which can cause non-linear distortion when large output signals are present is slew-rate limiting. The name refers to the fact that there is a specific maximum rate of change possible at the output of a real op-amp. This maximum is known as the slew rate (SR) of the

op-amp and is defined as:[34]

$$\text{SR} = \left.\frac{dv_o}{dt}\right|_{\text{max}} \tag{1.59}$$

and is usually specified on the op-amp data sheet in units of $\text{V} \cdot \text{ts}^{-1}$. It follows, if the input signal applied to an op-amp circuit is such that it demands an output response that is faster than the specified value of SR, the op-amp will not comply.

[34] Slew derives from a nautical term, and you can find many 1800s references to slewing a cannon on its trunnions (that would be a vertical movement). The earliest recorded usage (according to the OED) is from Falconer's Marine dictionary (1769):

> Rather, its output will change at the maximum possible rate, which is equal to its SR.

As an example, consider an op amp connected in the unity-gain voltage-follower configuration shown in Fig. 2.43(a), and let the input signal be the step voltage shown in Fig. 2.43(b). The output of the op amp will not be able to rise instantaneously to the ideal value V; rather, the output will be the linear ramp of slope equal to SR, shown in Fig. 2.43(c). The amplifier is then said to be **slewing**, and its output is slew-rate limited.

To understand the origin of the slew-rate phenomenon, we need to know about the internal circuit of the op-amp. For the time being, however, it is sufficient to know about the phenomenon and to note that it is distinct from the finite op-amp bandwidth that limits the frequency response of the closed-loop amplifiers, studied previously.

The limited bandwidth is a linear phenomenon and does not result in a change in the shape of an input sinusoid; that is, it does not lead to nonlinear distortion. The slew-rate limitation, on the other hand, can cause nonlinear distortion to an input sinusoidal signal when its frequency and amplitude are such that the corresponding ideal output would require vo to change at a rate greater than SR. This is the origin of another related op-amp specification, its full-power bandwidth, to be explained later.

Before leaving the example in Fig. 2.43, however, we should point out that if the step input voltage V is sufficiently small, the output can be the exponentially rising ramp shown in Fig. 2.43(d). Such an output would be expected from the follower if the only limitation on its dynamic performance were the finite op-amp bandwidth. Specifically, the transfer function of the follower can be found by substituting $R_1 = \infty$ and $R_2 = 0$ in Eq. (2.55) to obtain

$$\frac{V_o}{V_i} = \frac{1}{1 + s/\omega_i} \tag{1.60}$$

which is a low-pass STC response with a time constant $1/\omega_t$. Its step response would therefore be (see Appendix E)

$$v_o(t) = V(1 - e^{-\omega_t t}) \tag{1.61}$$

The initial slope of this exponentially rising function is $(\omega_t V)$. Thus, as long as V is sufficiently small so that $\omega_t V \leq SR$, the output will be as in Fig. 2.43(d).

1.9.4 Full-Power Bandwidth

Op-amp slew-rate limiting can cause non-linear distortion in sinusoidal waveforms. Consider once more the unity-gain follower with a sine-wave input given by:

$$v_I = \hat{V}_i \sin \omega t$$

The rate of change of this waveform is given as:

$$\frac{dv_I}{dt} = \omega \hat{V}_i \cos \omega t$$

with a maximum value of $\omega \hat{V}_i$. This maximum occurs at the zero crossings of the input sinusoid. Now if ωV_i exceeds the slew rate of the op amp, the output waveform will be distorted in the manner shown in Fig. 2.44. Observe that the output cannot keep up with the large rate of change of the sinusoid at its zero crossings, and the op amp slews.

The op-amp data sheets usually specify a frequency f_M called the **full-power bandwidth**. It is the frequency at which an output sinusoid with amplitude equal to the rated output voltage of the op-amp begins to show distortion due to slew-rate limiting. If we denote the rated output voltage V_{max}, then f_M is related to SR as follows:

$$\omega_M V_{o,max} = \text{SR} \qquad \text{and} \qquad f_M = \frac{\text{SR}}{2\pi V_{o,max}}$$

Based on the above given equation, it should be apparent the output sinusoids of amplitudes smaller than V_{max} will show slew-rate distortion at frequencies higher than ω_M.

In fact, at a frequency ω higher than ω_M, the maximum amplitude of the undistorted output sinusoid is given by

$$V_o = V_{o,max} = \left(\frac{\omega_M}{\omega}\right) \tag{1.62}$$

Chapter 2

Semiconductors

Table of Contents

2.1 Introduction 41
2.2 Intrinsic Semiconductors 43
2.3 Doped Semiconductors 46
2.4 Current Flow in Semiconductors 48
2.5 The PN Junction 52
2.6 PN Junction with Applied Voltage 57
2.7 Capacitive Effects in the pn Junction 61

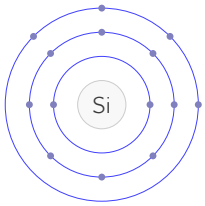
2.1 Introduction

In this chapter, we will have a gentle introduction the properties and physics of semiconductors. The objective is to provide a basis for understanding the physical operation of **diodes** and **transistors** to enable their effective use in the design of circuits. While many concepts studied in this chapter apply to semiconductor materials in general, our focus will be heavily biased towards **silicon**, as it is the primary material used in the vast majority of microelectronic circuits.[1]

This chapter will begins with a study of the crystal structure of semiconductors and introduces the two (2) types of charge carriers available for current conduction:

electrons, and holes.

> The most significant property of semiconductors is their conductivity can be varied over a very wide range through the introduction of controlled amounts of impurity atoms into the semiconductor crystal in a process called **doping**.



[1] The atomic structure and the free electrons present in a neutral Silicon atom.

Silicon is around 99% of semiconductor production volume and revenue as of 2025. Compound semiconductor (GaAs, InP, AlGaAs, SiC, SiGe, etc.) account for 1% combined.

Now we shall start to look at the building blocks which make up a semiconductor.

2.2 Intrinsic Semiconductors

As their name gives it away, semiconductors are materials whose conductivity lies between a of conductors, such as copper, and insulators, such as glass.

There are two (2) kinds of semiconductors:

- single-element semiconductors, such as germanium and silicon,[2] which are in group IV in the periodic table; and
- compound semiconductors, such as gallium-arsenide, which are formed by combining elements from groups III and V or groups II and VI.

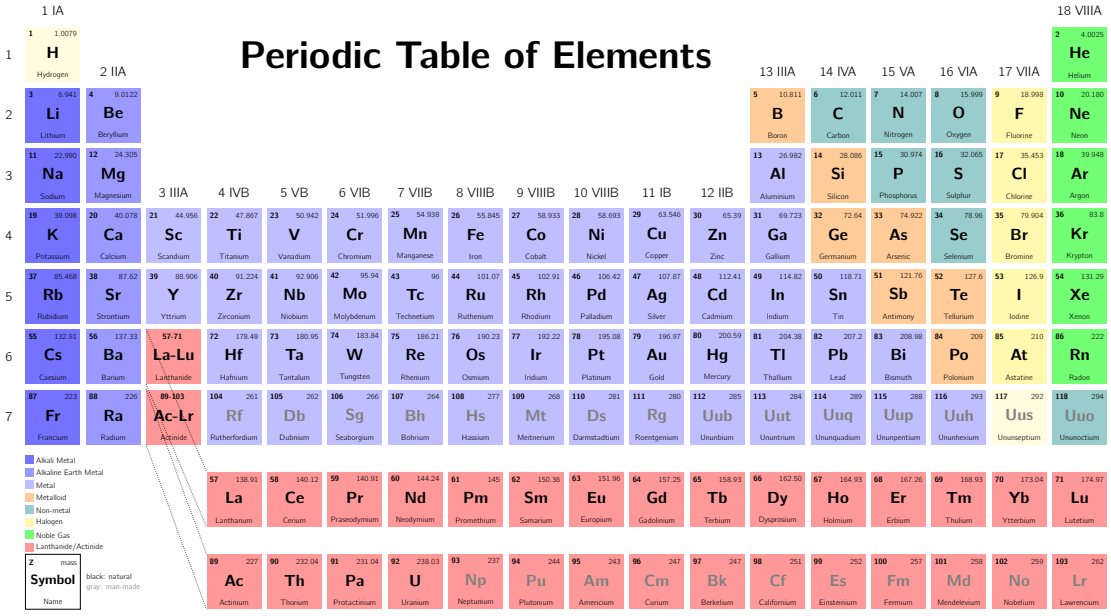


Figure 2.1: The periodic table of elements.

Compound semiconductors are useful in special electronic circuit applications as well as in applications that involve light, such as Light-Emitting Diode (LED)s. Of the two (2) elemental semiconductors, germanium[3] was used in the fabrication of very early transistors (late 1940s, early 1950s). It was quickly supplanted, however, with silicon, on which today's IC technology is almost entirely based.

> For this reason, we will deal mostly with silicon devices throughout this lecture series.

A silicon atom has four (4) valence electrons, and requires another four (4) to complete its outermost shell. This is achieved by sharing one of its valence electrons with each of its four neighboring atoms. Each pair of shared electrons forms a **covalent bond**. The result is a crystal of pure or intrinsic silicon with a regular lattice structure, where the atoms are held in their position by the covalent bonds.

[2] Silicon is an element; it has symbol Si and atomic number 14. It is a hard, brittle crystalline solid with a blue-grey metallic lustre, and is a tetravalent metalloid (sometimes considered as a non-metal) and semiconductor. It is a member of group 14 in the periodic table: carbon is above it; and germanium, tin, lead, and flerovium are below it. It is relatively unreactive. Silicon is a significant element that is essential for several physiological and metabolic processes in plants. Silicon is widely regarded as the predominant semiconductor material due to its versatile applications in various electrical devices such as transistors, solar cells, integrated circuits, and others.

[3] Germanium is a chemical element; it has symbol Ge and atomic number 32. It is lustrous, hard-brittle, grayish-white and similar in appearance to silicon. It is a metalloid or a nonmetal in the carbon group that is chemically similar to silicon. Like silicon, germanium naturally reacts and forms complexes with oxygen in nature.

Elemental germanium is used as a semiconductor in transistors and various other electronic devices. Historically, the first decade of semiconductor electronics was based entirely on germanium.

At sufficiently low temperatures,[4] all the covalent bonds are intact and no electrons are available to conduct electric current.

[4] close to, if not, absolute zero (0 K.)

At such low temperatures, the intrinsic silicon crystal behaves as an insulator. At room temperature, sufficient thermal energy exists to break some of the covalent bonds, a process known as thermal generation.

As shown in Fig. 3.2, when a covalent bond is broken, an electron is freed. The free electron can wander away from its parent atom, and it becomes available to conduct electric current if an electric field ($\boldsymbol{E}$) is applied to the crystal. As the electron leaves its parent atom, it leaves behind a net **positive charge** which is equal to the magnitude of the electron charge.

Therefore, an electron from a neighbouring atom could be attracted to this positive charge, and leaves its parent atom. This action fills up the "hole" which existed in the ionized atom but creates a new hole in the other atom.

This process may repeat itself, with the result that we effectively have a positively charged carrier, or **hole**, moving through the silicon crystal structure and being available to conduct electric current. The charge of a hole is equal in magnitude to the charge of an electron. We can therefore see that, as temperature increases, more covalent bonds are broken and electron-hole pairs are generated.

The increase in the numbers of free electrons and holes results in an increase in the conductivity of silicon.

Thermal generation results in free electrons and holes in equal numbers and hence equal concentrations, where concentration refers to the number of charge carriers per unit volume (cm^3). The free electrons and holes move randomly through the silicon crystal structure, and in the process some electrons may fill some of the holes. This process, called **recombination**, results in the disappearance of free electrons and holes. The recombination rate is proportional to the number of free electrons and holes, which in turn is determined by the thermal generation rate. The latter is a strong function of temperature.

In thermal equilibrium, the recombination rate is equal to the generation rate, and we can conclude that the concentration of free electrons (n) is equal to the concentration of holes (p),

$$n = p = n_i \tag{2.1}$$

where n_i denotes the number of free electrons and holes in a unit volume (cm^3) of intrinsic silicon at a given temperature. Results from semiconductor physics gives n_i as

$$n_i = BT^{3/2} \exp -E_g/2kT \tag{2.2}$$

where B is a material-dependent parameter which is $7{,}3 \times 10^{15}$ for silicon, T is the temperature in K; E_g, a parameter known as the **bandgap energy**, is 1.12 electron volt[5] (eV) for silicon; and k is Boltzmann's constant (8.62×10^{-5} eV/K).

[5] also written electron-volt and electron volt, is the measure of an amount of kinetic energy gained by a single electron accelerating through an electric potential difference of one volt in vacuum. When used as a unit of energy, the numerical value of $1\,e \cdot V$ in joules (J) is equal to the numerical value of the charge of an electron in coulombs (C).

Under the 2019 revision of the SI, this sets $1\,e \cdot V$ equal to the exact value $1{,}602{,}176{,}634 \times 10^{-19}$

The bandgap energy (E_g) is the minimum energy required to break a covalent bond and generate an electron-hole pair and is defined as an energy range in a solid where no electronic states exist.

Finally, it is useful for future purposes to express the product of the hole and free-electron concentration as

$$pn = n_i^2 \tag{2.3}$$

where for silicon at room temperature, $n_i \approx 1{,}5 \times 10^{10}\,\mathrm{cm}^{-3}$. As will be seen shortly, this relationship extends to extrinsic or doped silicon as well.

proportional to the number of free electrons and holes, which in turn is determined by the thermal generation rate. The latter is a strong function of temperature. In thermal equilibrium, the recombination rate is equal to the generation rate, and one can conclude that the concentration of free electrons n is equal to the concentration of holes p,

Information: Liquid Crystal Display

The existence of liquid crystals whose color could be changed by means of an external heat source was first reported in 1888 by an Austrian botanical physiologist. The LC idea lay dormant until the late 1940s, however. Subsequent developments in the field of solid-state electronics provided the technology to harness the technique in display media, with the first LCDs being demonstrated by RCA beginning in 1962. Today, LCDs are an essential component in every mobile device as the interface to the world of electronics within. At the other end of the scale, large LCDs are used in flat-panel TVs, and very large LCDs are appearing as “dynamic” wallpaper in museum display settings.

2.3 Doped Semiconductors

The intrinsic silicon crystal described previously has equal concentrations of free electrons and holes, generated by thermal generation. These concentrations are far too small for silicon to conduct measurable current at room temperature. In addition, the carrier concentrations and hence the conductivity are strong functions of temperature, which is **NOT** a desirable property in an electronic device. Fortunately, a method was developed to change the carrier concentration in a semiconductor crystal substantially and in a precisely controlled manner.

> This process is known as **doping**, and the resulting silicon is referred to as doped silicon.

Doping requires introducing impurity atoms into the silicon crystal in sufficient numbers to significantly increase the concentration of either free electrons or holes but with little or no change in the crystal properties of silicon.

To increase the concentration of free electrons (n) silicon is doped with an element with a valence of 5, such as *phosphorus*. The resulting doped silicon is then said to be of n type.[6] To increase the concentration of holes (p) silicon is doped with an element having a valence of 3, such as *boron*, and the resulting doped silicon is said to be of p type.

[6] This means, again, that the material has more free moving electrons than there are holes.

Figure 3.3 shows a silicon crystal doped with phosphorus impurity. The dopant (phosphorus) atoms replace some of the silicon atoms in the crystal structure. Since the phosphorus atom has five (5) electrons in its outer shell, four of these electrons form covalent bonds with the neighboring atoms, and the fifth electron becomes a free electron. Thus each phosphorus atom donates a free electron to the silicon crystal, and the phosphorus impurity is called a donor. It should be clear, though, that no holes are generated by this process. The net positive charge associated with the phosphorus atom is a bound charge that does not move through the crystal.

If the concentration of donor atoms is N_D, where N_D is usually much greater than n_i, the concentration of free electrons in the n-type silicon will be:

$$n_n \approx N_D \tag{2.4}$$

where the subscript n denotes n-type silicon. Therefore, n_n is determined by the doping concentration and not by temperature. This is not the case, however, for the hole concentration. All the holes in the n-type silicon are those generated by thermal ionization. Their concentration p_n can be found by noting that the relationship in Eq. (**2.3**) applies equally well for doped silicon, provided thermal equilibrium is achieved. Therefore for n-type silicon:

$$p_n n_n = n_i^2$$

Substituting for n_n from Eq. (**2.4**), we obtain for p_n:

$$p_n \simeq \frac{n_i^2}{N_n} \tag{2.5}$$

Therefore p_n will have the same dependence on temperature as that of n_i^2. Finally, we note that in n-type silicon the concentration of free electrons n_n will be much larger than that of holes. Hence electrons are said to be the majority charge carriers and holes the minority charge carriers in n-type silicon.

To obtain p-type silicon in which holes are the majority charge carriers, a trivalent impurity such as boron is used. Figure 3.4 shows a silicon crystal doped with boron. Note that the boron atoms replace some of the silicon atoms in the silicon crystal structure. Since each boron atom has three electrons in its outer shell, it accepts an electron from a neighbouring atom, thus forming covalent bonds. The result is a hole in the neighboring atom and a bound negative charge at the **acceptor** (boron) atom. It follows that each acceptor atom provides a hole. If the acceptor doping concentration is N_A, where $N_A \gg n_i$, the hole concentration becomes

$$p_p \simeq N_A \tag{2.6}$$

where the subscript p denotes p-type silicon. Therefore, here the majority carriers are holes and their concentration is determined by N_A. The concentration of minority electrons can be found by using the relationship $p_p n_p = n_i^2$ and substituting for p_p from Eq. (**2.6**):

$$n_p \simeq \frac{n_i^2}{N_A} \tag{2.7}$$

Therefore, the concentration of the minority electrons will have the same temperature dependence as that of n_i^2.

It should be emphasized that a piece of n-type or p-type silicon is electrically neutral; the charge of the majority free carriers[7] are neutralized by the bound charges associated with the impurity atoms.

[7] electrons in the n-type and holes in the p-type silicon

2.4 Current Flow in Semiconductors

There are two (2) different mechanisms for the movement of charge carriers and therefore for current flow in semiconductors:

- drift, and
- diffusion.

Let's have a look at both these options in more detail and understand their effect on semiconductors.

2.4.1 Drift Current

When an electrical field (E) is established in a semiconductor crystal, holes are accelerated in the direction of E, and free electrons are accelerated in the direction **opposite** to that of E. This situation is illustrated in Fig. 3.5. The holes acquire a velocity $v_{\text{p-drift}}$ given by:

$$v_{\text{p-drift}} = \mu_{\text{p}} E \tag{2.8}$$

where μ_{n} is a constant called the **hole mobility** where it represents the degree of ease by which holes move through the silicon crystal in response to the electrical field E.

Given the velocity has the units of $\text{cm} \cdot \text{s}^{-1}$ and $\boldsymbol{E}$ has the units of $\text{V} \cdot \text{m}^{-1}$, we see from Eq. (**2.8**), the mobility (μ_{p}) needs to have the units of $\text{cm}^2 \cdot \text{V}^{-1} \cdot \text{s}^{-1}$.

For intrinsic[8] silicon:

$$\mu_{\text{p}} = 480\,\text{cm}^2 \cdot \text{V}^{-1} \cdot \text{s}^{-1}.$$

[8] chemically pure silicon with no added impurities, meaning its electrical conductivity comes only from thermally generated electrons and holes

The free electrons acquire a drift velocity ($v_{\text{n-drift}}$) given by:

$$v_{n\text{-drift}} = -\mu_n E \tag{2.9}$$

where the result is negative because the electrons move in the direction opposite to E. Here μ_n is the **electron mobility**.[9] Note that μ_n is about 2.5 times μ_n, signifying that electrons move with much greater ease through the silicon crystal than do holes.

[9] for intrinsic silicon is about $1{,}350 \times 10^3\,\text{cm}^2 \cdot \text{V}^{-1} \cdot \text{s}^{-1}$

Let's now return to the single-crystal silicon bar shown in Fig. 3.5. Let the concentration of holes be p and that of free electrons n. We wish to calculate the current component due to the flow of holes. Consider a plane perpendicular to the x direction. In one second, the hole charge that crosses that plane will be ($Aqpv_{p\text{-drift}}$) coulombs, where A is the cross-sectional area of the silicon bar and q is the magnitude of electron charge. This then must be the hole component of the drift current flowing through the bar,

$$I_p = Aqpv_{p\text{-drift}} \tag{2.10}$$

Substituting for $v_{p\text{-drift}}$ from Eq. (**2.8**), we obtain $I_p = Aqp\mu_p E$ We are usually interested in the current density (J_p), which is the current per unit cross- sectional area ($A \cdot m^{-2}$),

$$J_p = \frac{I_p}{A} = qp\mu_p E \qquad [A \cdot m^{-2}] \quad (\mathbf{2.11})$$

The current component due to the drift of free electrons can be found in a similar manner. Note, however, that electrons drifting from right to left result in a current component from left to right.

> This is because of the convention of taking the direction of current flow as the direction of flow of positive charge and opposite to the direction of flow of negative charge.

Thus, $I_n = -Aqnv_{n\text{-drift}}$ Substituting for $v_{n\text{-drift}}$ from Eq. (**2.9**), we obtain the current density $J_n = I_n/A$ as

$$J_n = qn\mu_n E \qquad [A \cdot m^{-2}] \quad (\mathbf{2.12})$$

The total drift current density can now be found by summing J_p and J_n from Eqs. (3.11) and (3.12),

$$J = J_n + J_n = q(p\mu_n + n\mu_n)E \qquad [A \cdot m^{-2}] \quad (\mathbf{2.13})$$

This relationship can be written as

$$J = \sigma E \qquad \text{or} \qquad J = E/\rho \qquad [A \cdot m^{-2}] \quad (\mathbf{2.14})$$

where the conductivity (σ) is given as:

$$\sigma = q(p\mu_p + n\mu_n) \qquad [\Omega^{-1}] \quad (\mathbf{2.15})$$

and the resistivity (ρ) is given by:

$$\rho \equiv \frac{1}{\sigma} = \frac{1}{q\left(p\mu_p + n\mu_n\right)} \qquad [\Omega]$$

Please observe Eq. (**2.14**) is a form of Ohm's law and can be written alternately as

$$\rho = \frac{E}{I} \qquad [A/m^2] \quad (\mathbf{2.16})$$

Therefore the units of ρ are obtained from:

$$\frac{V \cdot cm^{-1}}{A \cdot cm^{-2}} = \Omega \cdot cm$$

2.4.2 Diffusion Current

Carrier diffusion occurs when the density of charge carriers in a piece of semiconductor is **NOT** uniform.

> For example, if by some mechanism the concentration of, say, holes, is made higher in one part of a piece of silicon than in another, then holes will diffuse from the region of high concentration to the region of low concentration.[10]

The diffusion of charge carriers gives rise to a net flow of charge, or diffusion current.

[10] Such a diffusion process is like that observed if one drops a few ink drops in a water-filled tank.

As an example, consider the bar of silicon shown in **Fig.** 2.2a: By some unspecified process, we have arranged to inject holes into its left side. This continuous hole injection gives rise to and maintains a hole **concentration profile** such as that shown in **Fig.** 2.2b. This profile in turn causes holes to diffuse from left to right along the silicon bar, resulting in a hole current in the x direction. The magnitude of the current at any point is proportional to the slope of the concentration profile, or the concentration gradient, at that point,

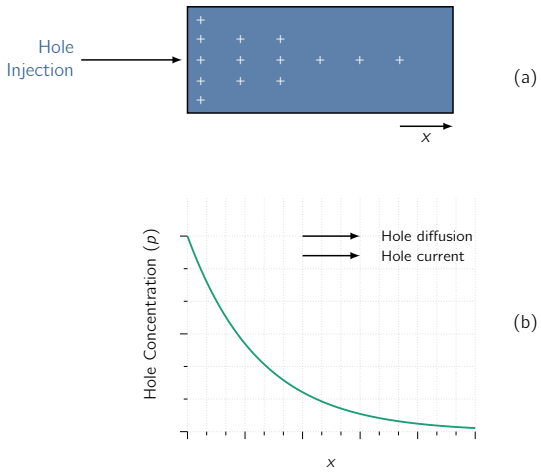
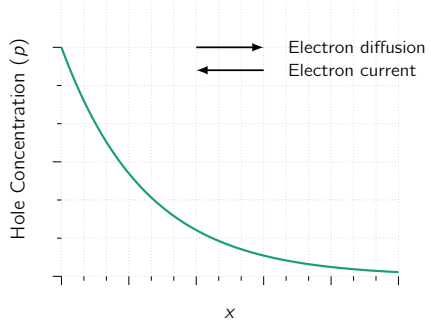


Figure 2.2 **A bar of silicon (a) into which holes are injected, thus creating the hole concentration profile along the x axis, shown in (b). The holes diffuse in the positive direction of x and give rise to a hole diffusion current in the same direction. Note that we are not showing the circuit to which the silicon bar is connected.**

$$J_p = -qD_p \frac{dp(x)}{dx} \tag{2.17}$$

where J_p is the hole-current density ($A \cdot cm^{-2}$), q is the magnitude of electron charge, D_p is a constant called the diffusion constant or diffusivity of holes; and $p(x)$ is the hole concentration at point x.

> The gradient dp/dx is negative, resulting in a positive current in the x direction, as should be expected.



In the case of electron diffusion resulting from an electron concentration gradient, as shown in **Fig.** 2.3, a similar relationship applies, giving the electron-current density,

$$J_n = qD_n \frac{dn(x)}{dx} \tag{2.18}$$

where D_n is the diffusion constant or diffusivity of electrons. Observe that a negative (dn/dx) gives rise to a negative current, a result of the convention that the positive direction of current is taken to be that of the flow of positive charge.[11]

Figure 2.3 **If the electron concentration profile shown is established in a bar of silicon, electrons diffuse in the x direction, giving rise to an electron diffusion current in the negative-x direction.**

[11] and opposite to that of the flow of negative charge

> For holes and electrons diffusing in intrinsic silicon, typical values for the diffusion constants are $D_p = 12\,cm^2 \cdot s^{-1}$ and $D_n = 35\,cm^2 \cdot s^{-1}$.

At this point the reader is probably wondering where the diffusion current in the silicon bar in Fig. 3.6(a) goes. A good question, as we are not showing how the right-side end of the bar is connected to the rest of the circuit. We will address this and related questions in detail in our discussion of the pn junction in later sections.

Relationship between D and μ

A simple yet powerful relationship ties the diffusion constant with the mobility:

$$\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} = V_T \qquad \text{where} \qquad V_T = \frac{kT}{q} \tag{2.19}$$

The parameter V_T is known as the **thermal voltage**. At room temperature, $T \simeq 300\,\text{K}$ and $V_T = 25{,}9\,\text{mV}$.

This relationship is known as the Einstein relationship.[12]

[12] In physics (specifically, the kinetic theory of gases), the Einstein relation is a previously unexpected[clarification needed] connection revealed independently by William Sutherland in 1904,[1][2][3] Albert Einstein in 1905,[4] and by Marian Smoluchowski in 1906[5] in their works on Brownian motion. The more general form of the equation in the classical case is[6]

2.5 The PN Junction

Having learned important semiconductor concepts, we are now ready to consider our first practical semiconductor structure:

the *pn* junction.

As previously mentioned, the *pn* junction implements the diode and plays the dominant role in the structure and operation of the Bipolar Junction Transistor (BJT).

> Understanding *pn* junctions is very important to the study of the MOSFET operation.

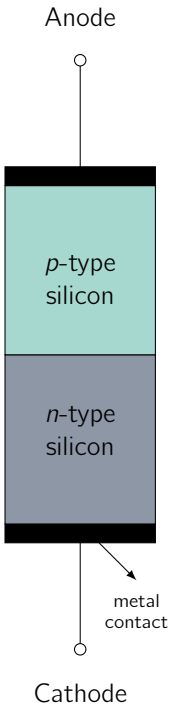


Figure 2.4 **Simplified physical structure of the pn junction. As the pn junction implements the junction diode, its terminals are labeled anode and cathode.**

2.5.1 Implementation

Fig. 2.4 shows a simplified physical structure of the *pn* junction. It consists of a *p*-type semiconductor[13] brought into close contact with an *n*-type semiconductor material, which is also silicon.

[13] e.g., silicon

> In actual practice, both *p* and *n* regions are part of the same silicon crystal. That is, the *pn* junction is formed within a single silicon crystal by creating regions of different dopings (*p* and *n* regions).

As indicated in **Fig.** 2.4, external wire connections are made to the *p* and *n* regions through metal (aluminum) contacts. If the *pn* junction is used as a diode, these constitute the diode terminals and are therefore labeled anode and cathode in keeping with diode terminology.[14]

[14] This terminology in fact is a carryover from that used with vacuum-tube technology, which was the technology for making diodes and other electronic devices until the invention of the transistor in 1947. This event ushered in the era of solid-state electronics, which changed not only electronics, communications, and computers but indeed the world!

2.5.2 Open Circuit Operation

Fig. 2.5 shows a *pn* junction under open-circuit conditions-that is, the external terminals are left open. The "+" signs in the *p*-type material denote the **majority holes**. The charge of these holes is neutralized by an equal amount of bound negative charge associated with the acceptor atoms.

For simplicity, these bound charges are not shown in the diagram. Also not shown are the minority electrons generated in the *p*-type material by thermal ionization. In the *n*-type material the majority electrons are indicated by "-" signs. Here also, the bound positive charge, which neutralizes the charge of the majority electrons, is not shown in order to keep the diagram simple. The *n*-type material also contains minority holes generated by thermal ionization but not shown in the diagram.

Diffusion Current

Because the concentration of holes is high in the *p* region and low in the *n* region, holes diffuse across the junction from the *p* side to the *n* side. Similarly, electrons diffuse across the junction from the *n* side to the *p* side.

These two current components add together to form the diffusion current I_p, whose direction is from the p side to the n side, as indicated in **Fig.** 2.5.

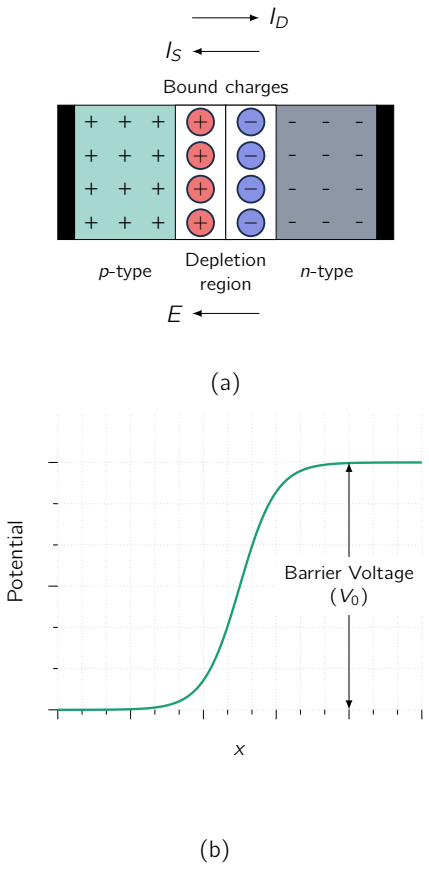


Figure 2.5
(a) The pn junction with no applied voltage (open-circuited terminals). (b) The potential distribution along an axis perpendicular to the junction.

Depletion Region

The holes that diffuse across the junction into the *n* region quickly recombine with some of the majority electrons present there and therefore disappear from the scene.

This recombination process results also in the disappearance of some free electrons from the *n*-type material. Therefore some of the bound positive charge will no longer be neutralized by free electrons, and this charge is said to have been uncovered. Given the recombination takes place close to the junction, there will be a region close to the junction that is depleted of free electrons and contains uncovered bound positive charge, as indicated in Fig. 3.9.

The electrons that diffuse across the junction into the *p* region quickly recombine with some of the majority holes there, and thus disappear from the scene. This results also in the disappearance of some majority holes, causing

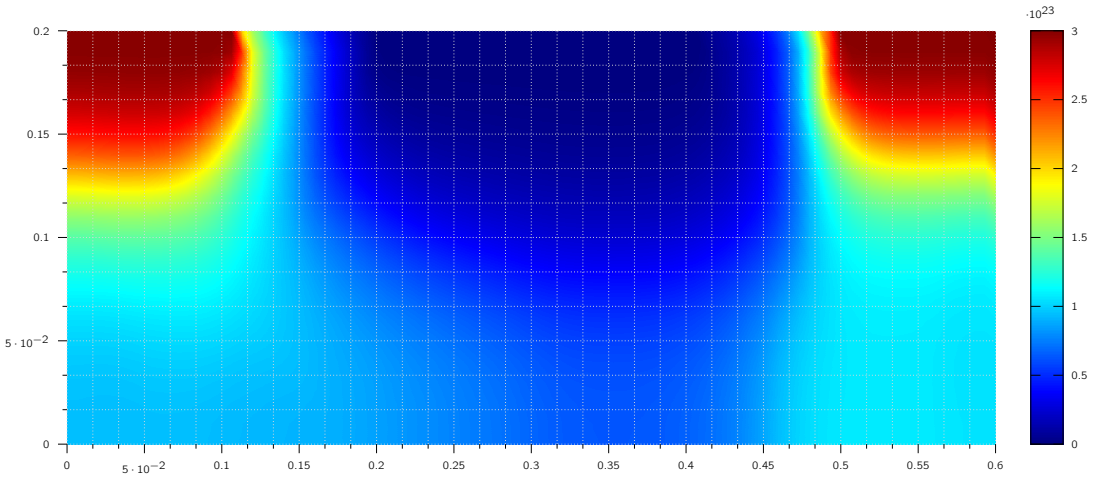


Figure 2.6
The electron density of a MESFET. Please observe the central point where the environment is void of electrons.

some of the bound negative charge to be uncovered (i.e., no longer neutralized by holes). Thus, in the *p* material close to the junction, there will be a region depleted of holes and containing uncovered bound negative charge, as indicated in **Fig.** 2.5.

From the above it follows that a **carrier-depletion region** will exist on both sides of the junction, with the *n* side of this region positively charged and the *p* side negatively charged. This carrier-depletion region-or, simply, depletion region-is also called the space-charge **region**. The charges on both sides of the depletion region cause an electric field *E* to be established across the region in the direction indicated in **Fig.** 2.5. Hence a potential difference results across the depletion region, with the *n* side at a positive voltage relative to the *p* side, as shown in **Fig.** 2.5b. Thus the resulting electric field opposes the diffusion of holes into the *n* region and electrons into the *p* region. In fact, the voltage drop across the depletion region acts as a **barrier** that has to be overcome for holes to diffuse into the *n* region and electrons to diffuse into the *p* region. The larger the barrier voltage, the smaller the number of carriers that will be able to overcome the barrier, and hence the lower the magnitude of diffusion current. Thus it is the appearance of the barrier voltage V_0 that limits the carrier diffusion process. It follows that the diffusion current *ID* depends strongly on the voltage drop *V*0 across the depletion region.

Drift Current and Equilibrium

In addition to the current component I_p due to majority-carrier diffusion, a component due to minority-carrier drift exists across the junction. Specifically, some of the thermally generated holes in the *n* material move toward the junction and reach the edge of the depletion region. There, they experience the electric field in the depletion region, which sweeps them across that region into the *p* side. Similarly, some of the minority thermally generated electrons in the *p* material move to the edge of the depletion region and get swept by the electric field in the depletion region across that region into the *n* side. These two current components-electrons moved by drift from *p* to *n* and holes moved by drift from n to p-add together to form the drift current I_s, whose direction is from the n side to the p side of the junction, as indicated in Fig. 3.9. Since the current I_s is carried by thermally generated minority carriers, its value is strongly dependent on temperature; however, it is independent of the value of the depletion-layer voltage V_0. This is due to the fact that the drift current is determined by the number of minority carriers that make it to the edge of the depletion region; any minority carriers that manage to get to the edge of the depletion region will be swept across by E irrespective of the value of E or, correspondingly, of V_0.

Under open-circuit conditions (**Fig.** 2.5) no external current exists; thus the two opposite currents across the junction must be equal in magnitude:

$$I_D = I_S$$

This equilibrium condition[15] is maintained by the barrier voltage (V_0). Therefore, if for some reason I_D exceeds I_S, then more bound charge will be uncovered on both sides of the junction, the depletion

[15] In fact, in equilibrum the equality of drift and diffusion currents applies not just to the total currents but also to their individual components. That is, the hole drift current must equal the hole diffusion current and, similarly, the electron drift current must equal the electron diffusion current.

layer will widen, and the voltage across it (V_0) will increase. This in turn causes I_D to decrease until equilibrium is achieved with $I_D = I_S$.

On the other hand, if I_S exceeds I_D, then the amount of uncovered charge will decrease, the depletion layer will narrow, and the voltage across it (V_0) will decrease. This causes I_D to increase until equilibrium is achieved with $I_D = I_S$.

The Junction Built-in Voltage

With no external voltage applied, the barrier voltage (V_0) across the *pn* junction can be shown to be given by:

$$V_0 = V_T \ln\left(\frac{N_A N_D}{n_i^2}\right) \tag{2.20}$$

where N_A and N_D are the doping concentrations of the *p*-side and *n*-side of the junction, respectively. Therefore V_0 depends both on doping concentrations and on temperature. It is known as the **junction built-in voltage**.

> Typically, for silicon at room temperature, V_0 is in the range of 0,6 V to 0,9 V.

When the *pn* junction terminals are left open-circuited, the voltage measured between them will be zero. That is, the voltage V_0 across the depletion region does **NOT** appear between the junction terminals. This is because of the contact voltages existing at the metal-semiconductor junctions at the terminals, which counter and exactly balance the barrier voltage. If this were not the case, we would have been able to draw energy from the isolated pn junction, which would clearly violate the principle of conservation of energy.

Width of and Charge Stored in the Depletion Region

Figure 3.10 provides further illustration of the situation occuring in the *pn* junction when the junction is in equilibrium. In Fig. 3.10(a) we see a junction in which $N_A > N_D$, a typical situation in practice. This is borne out by the carrier concentration on both sides of the junction, as shown in Fig. 3.10(b).

> We have denoted the minority-carrier concentrations in both sides by n_{p0} and p_{n0}, with the additional subscript "0" signifying equilibrium.[16]

Observe that the depletion region extends in both the *p* and *n* materials and equal amounts of charge exist on both sides.[17] However, since usually unequal dopings N_A and N_D are used, as in the case illustrated in Fig. 3.10, the width of the depletion layer will **NOT** be the same on the two (2) sides. Rather, to uncover the same amount of charge, the depletion layer will extend deeper into the more lightly doped material. Specifically, if we denote the width of the depletion region in the *p*

[16] i.e., before external voltages are applied, as will be seen in the next section.

[17] (Q_+ and Q_- in Fig. 3.10c)

side by x_n and in the n side by x_n, we can express the magnitude of the charge of the junction as

$$\text{on the } n \text{ side} \qquad |Q_+| = qAx_nN_D \qquad \text{and on the } p \text{ side} \qquad |Q_-| = qAx_pN_A \tag{2.21}$$

where A is the cross-sectional area of the junction in the plane perpendicular to the page. The charge equality condition can now be written as $qAx_nN_D = qAx_nN_A$ which can be rearranged to give us:

$$\frac{x_n}{r} = \frac{N_A}{N_B} \tag{2.22}$$

In actual practice, it is usual for one side of the junction to be much more heavily doped than the other, with the result that the depletion region exists almost entirely on one side (the lightly doped side).

The width W of the depletion layer can be shown to be given by

$$W = x_n + x_p = \sqrt{\frac{2\epsilon_s}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)V_0} \tag{2.23}$$

where ϵ_s is the electrical permittivity of silicon = $11.7\epsilon_0 = 11.7 \times 8.85 \times 10^{-14}$ F/cm = 1.04×10^{-12} F/cm. Typically W is in the range 0.1 μm to 1 μm. Eqs. (3.25) and (3.26) can be used to obtain x_n and x_p in terms of W as:

$$x_n = W\frac{N_A}{N_A + N_D} \qquad \text{and} \qquad x_p = W\frac{N_D}{N_A + N_D} \tag{2.24}$$

The charge stored on either side of the depletion region can be expressed in terms of W by utilizing Eqs. (3.23) and Eq. (**2.24**) to obtain:

$$Q_I = |Q_+| = |Q_-| \tag{2.25}$$

$$Q_J = Aq\left(\frac{N_AN_D}{N_A + N_D}\right)W \tag{2.26}$$

Finally, we can substitute for W from Eq. (3.26) to obtain

$$Q_J = A\sqrt{2\epsilon_s q\left(\frac{N_AN_D}{N_A + N_D}\right)V_0} \tag{2.27}$$

These expressions for QJ will prove useful in subsequent sections.

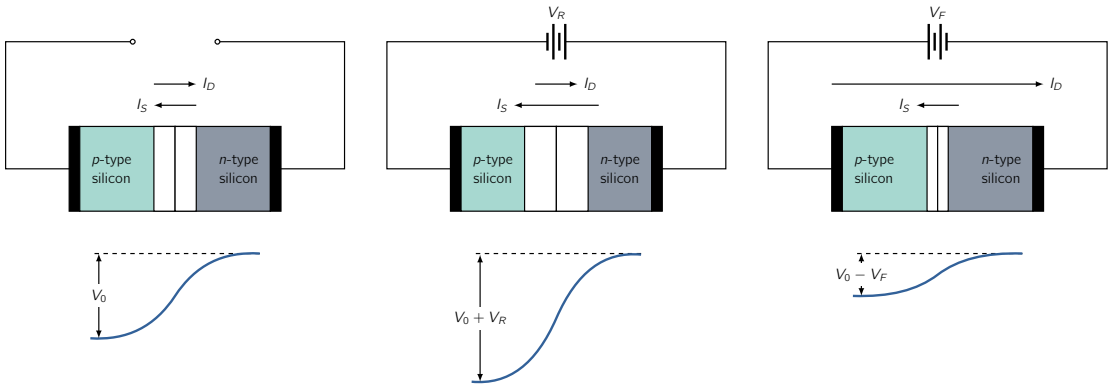


Figure 2.7

2.6 PN Junction with Applied Voltage

Now that we had a look at the open-circuited *pn* junction in detail, we are now ready to apply a DC voltage between its two (2) terminals to find its electrical conduction properties. If the voltage is applied so that the *p* side is made more positive than the *n* side, it is referred to as a forward-bias voltage.[18] Conversely, if our applied dc voltage is such that it makes the *n* side more positive than the *p* side, it is said to be a reverse-bias voltage. As will be seen, the *pn* junction exhibits vastly different conduction properties in its forward and reverse directions. Our plan is as follows. We begin by a simple qualitative description in Section 3.5.1 and then consider an analytical description of the $i - v$ characteristic of the junction in Section 3.5.2.

[18] we take the term bias to refer simply to the application of a dc voltage.

2.6.1 Qualitative Description of Junction Operation

Fig. 2.7 shows the *pn* junction under three different conditions: (a) the open-circuit or equilibrium condition studied in the previous section; (b) the reverse-bias condition, where a DC voltage V_R is applied; and (c) the forward-bias condition, where a DC voltage V_F is applied.

Observe that in the open-circuit case, a barrier voltage V_0 develops, making n more positive than p, and limiting the diffusion current Ip to a value exactly equal to the drift current Is, thus resulting in a zero current at the junction terminals, as should be the case, since the terminals are open-circuited. Also, as mentioned previously, the barrier voltage V_0, though it establishes the current equilibrium across the junction, does not in fact appear between the junction terminals.

Consider now the reverse-bias case in (b). The externally applied reverse-bias voltage V_R is in the direction to add to the barrier voltage, and it does, thus increasing the effective barrier voltage to $(V_0 + V_R)$ as shown. This reduces the number of holes that diffuse into the *n* region and the number of electrons that diffuse into the *p* region. The end result is that the diffusion current ID is dramatically reduced. As will be seen shortly, a reverse-bias voltage of a volt or so is sufficient to

cause $I_p \simeq 0$, and the current across the junction and through the external circuit will be equal to I_s. Recalling that I_s is the current due to the drift across the depletion region of the thermally generated minority carriers, we expect I_s to be very small and to be strongly dependent on temperature. We will show this to be the case very shortly. We thus conclude that in the reverse direction, the *pn* junction conducts a very small and almost-constant current equal to I_s.

Before leaving the reverse-bias case, observe that the increase in barrier voltage will be accompanied by a corresponding increase in the stored uncovered charge on both sides of the depletion region. This in turn means a wider depletion region, needed to uncover the additional charge required to support the larger barrier voltage $(V_0 + V_R)$. Analytically, these results can be obtained easily by a simple extension of the results of the equilibrium case. Thus the width of the depletion region can be obtained by replacing V_0 in Eq. (3.26) by $(V_0 + V_0)$,

$$W = x_n + x_p = \sqrt{\frac{2\epsilon_s}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)(V_0 + V_R)} \tag{2.28}$$

and the magnitude of the charge stored on either side of the depletion region can be determined by replacing V_0 in Eq. (3.30) by $(V_0 + V_R)$,

$$Q_J = A\sqrt{2\epsilon_s q\left(\frac{N_A N_D}{N_A + N_D}\right)(V_0 + V_R)} \tag{2.29}$$

We next consider the forward-bias case shown in Fig. 3.11(c). Here the applied voltage V_F is in the direction that subtracts from the built-in voltage V_0, resulting in a reduced barrier voltage $(V_0 - V_F)$ across the depletion region. This reduced barrier voltage will be accompanied by reduced depletion-region charge and correspondingly narrower depletion-region width W. Most importantly, the lowering of the barrier voltage will enable more holes to diffuse from p to n and more electrons to diffuse from n to p. Thus the diffusion current Ip increases substantially and, as will be seen shortly, can become many orders of magnitude larger than the drift current I_s. The current I in the external circuit is of course the difference between I_p and I_s, $I = I_p - I_s$ and it flows in the forward direction of the junction, from p to n. We thus conclude that the pn junction can conduct a substantial current in the forward-bias region and that current is mostly a diffusion current whose value is determined by the forward-bias voltage V_F.

Substituting for $p_n(x)$ from Eq. (3.35) gives

$$J_p(x) = q\left(\frac{D_p}{L_p}\right)p_{n0}\left(e^{V/V_T} - 1\right)e^{-(x-x_n)/L_p} \tag{2.30}$$

As expected, $J_p(x)$ is highest at $x = x_n$,

$$J_p(x_n) = q\left(\frac{D_p}{L_n}\right)p_{n0}\left(e^{V/V_T} - 1\right) \tag{2.31}$$

and decays exponentially for $x > x_n$, as the minority holes recombine with the majority electrons. This recombination, however, means that the majority electrons will have to be replenished by a current that injects electrons from the external circuit into the n region of the junction. This latter current component has the same direction as the hole current (because electrons moving from right

to left give rise to current in the direction from left to right). It follows that as $J_p(x)$ decreases, the electron current component increases by exactly the same amount, making the total current in the n material constant at the value given by Eq. (3.37).

An exactly parallel development can be applied to the electrons that are injected from the n to the p region, resulting in an electron diffusion current given by a simple adaptation of Eq. (3.37),

$$J_n(-x_p) = q\left(\frac{D_n}{L_n}\right) n_{p0}\left(e^{V/V_T} - 1\right) \tag{2.32}$$

Now, although the currents in Eqs. (3.37) and (3.38) are found at the two edges of the depletion region, their values do not change in the depletion region. Thus we can drop the location descriptors (x_n), $(-x_p)$, add the two current densities, and multiply by the junction area A to obtain the total current I as $I = A(J_p + J_n)$ $I = Aq\left(\frac{D_p}{L_n}p_{n0} + \frac{D_n}{L_n}n_{p0}\right)\left(e^{V/V_T} - 1\right)$ Substituting for $p_{n0} = n_i^2/N_D$ and for $n_{p0} = n_i^2/N_A$ gives

$$I = Aqn_i^2\left(\frac{D_p}{L_nN_D} + \frac{D_n}{L_nN_A}\right)\left(e^{V/V_T} - 1\right) \tag{2.33}$$

From this equation we note that for a negative V (reverse bias) with a magnitude of a few times V_T (25.9 mV), the exponential term becomes essentially zero, and the current across the junction becomes negative and constant. From our qualitative description in Section 3.5.1, we know that this current must be I_s. Thus,

$$I = I_s(e^{V/V_T} - 1) \tag{2.34}$$

where

$$I_S = Aqn_i^2\left(\frac{D_p}{L_pN_p} + \frac{D_n}{L_nN_A}\right) \tag{2.35}$$

Figure 3.13 shows the $I - V$ characteristic of the pn junction (Eq. 3.40). Observe that in the reverse direction the current saturates at a value equal to $-I_s$. For this reason, I_s is given the name **saturation current**. From Eq. (3.41) we see that I_s is directly proportional to the cross-sectional area A of the junction. Thus, another name for I_s, one we prefer to use in this book, is the junction scale current. Typical values for I_s, for junctions of various areas, range from 10^{-18} A to 10^{-12} A.

Besides being proportional to the junction area A, the expression for I_s in Eq. (3.41) indicates that I_s is proportional to n_i^2, which is a very strong function of temperature (see Eq. 3.2).

2.6.2 Reverse Breakdown

The description of the operation of the pn junction in the reverse direction, and the $I-V$ relationship of the junction in Eq. (**2.34**), indicate that at a reverse-bias voltage $(-V)$, with the condition of $V \gg Vr$, when applied will result in reverse current that flows across the junction is approximately equal to I_S and therefore is very small. However, as the magnitude of the reverse-bias voltage V is increased, a value is reached at which a very large reverse current flows as shown in Fig. 3.14.

Observe that as V reaches the value $V[Z]$, the dramatic increase in reverse current is accompanied by a very small increase in the reverse voltage; that is, the reverse voltage across the junction

remains very close to the value V_Z. The phenomenon that occurs at $V = V_Z$ is known as **junction breakdown**.

> It is **NOT** a destructive phenomenon. That is, the *pn* junction can be repeatedly operated in the breakdown region without a permanent effect on its characteristics.

This, however, is predicated on the assumption that the magnitude of the reverse-breakdown current is limited by the external circuit to a safe value. The "safe" value is one that results in the limitation of the power dissipated in the junction to a safe, allowable level.

There are two (2) possible mechanisms for *pn* junction breakdown:

- the zener effect,[19] and
- avalanche effect.

[19] Named after an early worker in the area. Note that the subscript Z in V_Z denotes zener. We will use V_Z to denote the breakdown voltage whether the breakdown mechanism is the zener effect or the avalanche effect.

If a pn junction breaks down with a breakdown voltage $V_Z < 5$ V, the breakdown mechanism is usually the zener effect. Avalanche breakdown occurs when V_Z is greater than approximately 7 V. For junctions that break down between 5 V and 7 V, the breakdown mechanism can be either the zener or the avalanche effect or a combination of the two.

Zener breakdown occurs when the electric field in the depletion layer increases to the point of breaking covalent bonds and generating electron-hole pairs. The electrons generated in this way will be swept by the electric field into the *n* side and the holes into the *p* side. Therefore these electrons and holes constitute a reverse current across the junction. Once the zener effect starts, a large number of carriers can be generated, with a negligible increase in the junction voltage. Therefore the reverse current in the breakdown region will be large and its value must be determined by the external circuit, while the reverse voltage appearing between the diode terminals will remain close to the specified breakdown voltage V_Z.

The other breakdown mechanism, avalanche breakdown, occurs when the minority carriers that cross the depletion region under the influence of the electric field gain sufficient kinetic energy to be able to break covalent bonds in atoms with which they collide. The carriers liberated by this process may have sufficiently high energy to be able to cause other carriers to be liberated in another ionizing collision. This process keeps repeating in the fashion of an avalanche, with the result that many carriers are created that are able to support any value of reverse current, as determined by the external circuit, with a negligible change in the voltage drop across the junction.

As will be seen when we study Diodes in the next chapter, some pn junction diodes are fabricated to operate specifically in the breakdown region, where use is made of the nearly constant voltage V_Z.

2.7 Capacitive Effects in the pn Junction

There are two (2) charge-storage mechanisms in a given *pn* junction.

> One is associated with the charge stored in the depletion region, and the other is associated with the minority-carrier charge stored in the *n* and *p* materials as a result of the concentration profiles established by carrier injection.

> While the first is easier to see when the *pn* junction is reverse biased, the second is in effect only when the junction is forward biased.

2.7.1 Depletion or Junction Capacitance

When a *pn* junction is reverse biased with a voltage V_R, the charge stored on either side of the depletion region is given by Eq. (**2.29**),

$$Q_J = A\sqrt{2\epsilon_s q\frac{N_A N_D}{N_A + N_D}}(V_0 + V_R)$$

Thus, for a given pn junction,

$$Q_j = \alpha\sqrt{V_0 + V_R} \qquad \text{where} \qquad \alpha = A\sqrt{2\epsilon_s q\frac{N_A N_D}{N_A + N_D}} \tag{2.36}$$

Thus Q_J is nonlinearly related to V_R, as shown in Fig. 3.15. This non-linear relationship makes it difficult to define a capacitance which accounts for the need to change Q_J whenever V_R is changed. We can, however, assume the junction is operating at a point such as Q, as indicated in Fig. 3.15, and define a capacitance C_j which relates the change in the charge Q_j to a change in the voltage V_R,

$$C_j = \left.\frac{dQ_j}{dV_R}\right|_{V_R - V_Q} \tag{2.37}$$

This incremental-capacitance approach turns out to be quite useful in electronic circuit design, as we shall see throughout this book. Using Eq. (**2.37**) together with Eq. (**2.36**) gives us:

$$C_j = \frac{\alpha}{2\sqrt{V_0 + V_p}} \tag{2.38}$$

The value of C_j at zero reverse bias can be obtained from Eq. (**2.38**) as

$$C_{j0} = \frac{\alpha}{2\sqrt{V_0}} \tag{2.39}$$

which enables us to express C_j as:

$$C_j = \frac{C_{j0}}{\sqrt{1 + \frac{V_R}{V_0}}} \tag{2.40}$$

where C_{j0} is given by Eq. (**2.39**) or alternatively if we substitute for α from Eq. (**2.36**) by

$$C_{j0} = A\sqrt{\left(\frac{\epsilon_s q}{2}\right)\left(\frac{N_A N_D}{N_A + N_D}\right)\left(\frac{1}{V_0}\right)} \tag{2.41}$$

Before leaving the subject of depletion-region or junction capacitance we point out that in the *pn* junction we have been studying, the doping concentration is made to change abruptly at the junction boundary. Such a junction is known as an **abrupt junction**. There is another type of pn junction in which the carrier concentration is made to change gradually from one side of the junction to the other. To allow for such a graded junction, the formula for the junction capacitance (Eq. (**2.40**)) can be written in the more general form

$$C_j = \frac{C_{j0}}{\left(1 + \frac{V_R}{V}\right)^m} \tag{2.42}$$

where m is a constant called the **grading coefficient**, whose value ranges from 1/3 to 1/2 depending on the manner in which the concentration changes from the p to the n side.

2.7.2 Diffusion Capacitance

Consider a forward-biased *pn* junction. In steady state, minority-carrier distributions in the p and n materials are established, as shown in Fig. 3.12. Thus a certain amount of excess minority-carrier charge is stored in each of the p and n bulk regions.[20] If the terminal voltage V changes, this charge will have to change before a new steady state is achieved. This charge-storage phenomenon gives rise to another capacitive effect, distinctly different from that due to charge storage in the depletion region.

[20] outside the depletion region.

To calculate the excess minority-carrier charge, refer to Fig. 3.12. The excess hole charge stored in the n region can be found from the shaded area under the exponential as follows:[21]

$$\begin{aligned} Q_n &= Aq \times \quad \text{shaded area under the } p_n(x) \text{ curve} \\ &= Aq[p_n(x_n) - p_{n0}]L_n \end{aligned}$$

[21] Recall that the area under an exponential curve $Ae^{-x/B}$ is equal to AB.

Substituting for $p_n(x_n)$ from Eq. (3.33) and using Eq. (**2.31**) enables us to express Q_p as

$$Q_p = \frac{L_p^2}{D} I_p \tag{2.43}$$

The factor (L_p^2/D_p) that relates Q_p to I_p is a useful device parameter that has the dimension of time (s) and is denoted τ_p

$$\tau_p = \frac{L_p^2}{D_p} \qquad \text{therefore} \qquad Q_p = \tau_p I_p \tag{2.44}$$

The time constant τ_p is known as the excess **minority-carrier** (hole) lifetime. It is the average time it takes for a hole injected into the n region to recombine with a majority electron. This definition of τ_p implies that the entire charge Q_p disappears and has to be replenished every τ_p seconds.

The current that accomplishes the replenishing is $I_p = Q_p/\tau_p$. This is an alternate derivation for Eq. (**2.44**).

A relationship similar to that in Eq. (**2.44**) can be developed for the electron charge stored in the p region,

$$Q_n = \tau_n I_n \tag{**2.45**}$$

where τ_n is the electron lifetime in the p region. The total excess minority-carrier charge can be obtained by adding together Q_p and Q_n,

$$Q = \tau_p I_p + \tau_n I_n \tag{**2.46**}$$

This charge can be expressed in terms of the diode current $I = I_p + I_n$ as

$$Q = \tau_r I \tag{**2.47**}$$

where τ_T is called the **mean transit time** of the junction. Obviously, τ_T is related to τ_p and τ_n. Furthermore, for most practical devices, one side of the junction is much more heavily doped than the other. For instance, if $N_A \gg N_D$, one can show that $I_p \gg I_n$, $I \simeq I_p$, $Q_p \gg Q_n$, $Q \simeq Q_p$, and thus $\tau_T \simeq \tau_p$.

For small changes around a bias point, we can define an incremental diffusion capacitance C_d as

$$C_\mathrm{d} = \frac{dQ}{dV} \tag{**2.48**}$$

and can show that

$$C_\mathrm{d} = \left(\frac{\tau_\mathrm{T}}{V_\mathrm{t}}\right) I \tag{**2.49**}$$

where I is the forward-bias current.

> As can be seen C_d is directly proportional to the forward current (I) and therefore is negligibly small when the diode is reverse biased. In addition, to keep C_d small, the transit time τ_T **must** be made small, an important requirement for a pn junction intended for high-speed or high-frequency operation.

Chapter Symbols and Units

Symbol	Description	Unit

Part II

Semiconductor Devices

Index

bandgap energy, 44

closed-loop gain, 14
common-mode rejection, 6

diffusion current, 49
donor, 46
doping, 46

Op-Amp
    closed-loop gain, 15
    closed-loop-gain, 15
    degenerative feedback, 15
    differential gain, 7
    negative feedback, 9
    non-inverting amplifier, 15
    open-loop gain, 15
    positive feedback, 9
    virtual short circuit, 10, 14
    voltage follower, 16
open-loop gain, 7

recombination, 44, 45

single-time-constant, 34

thermal generation, 44–46

unity-gain-bandwidth, 34