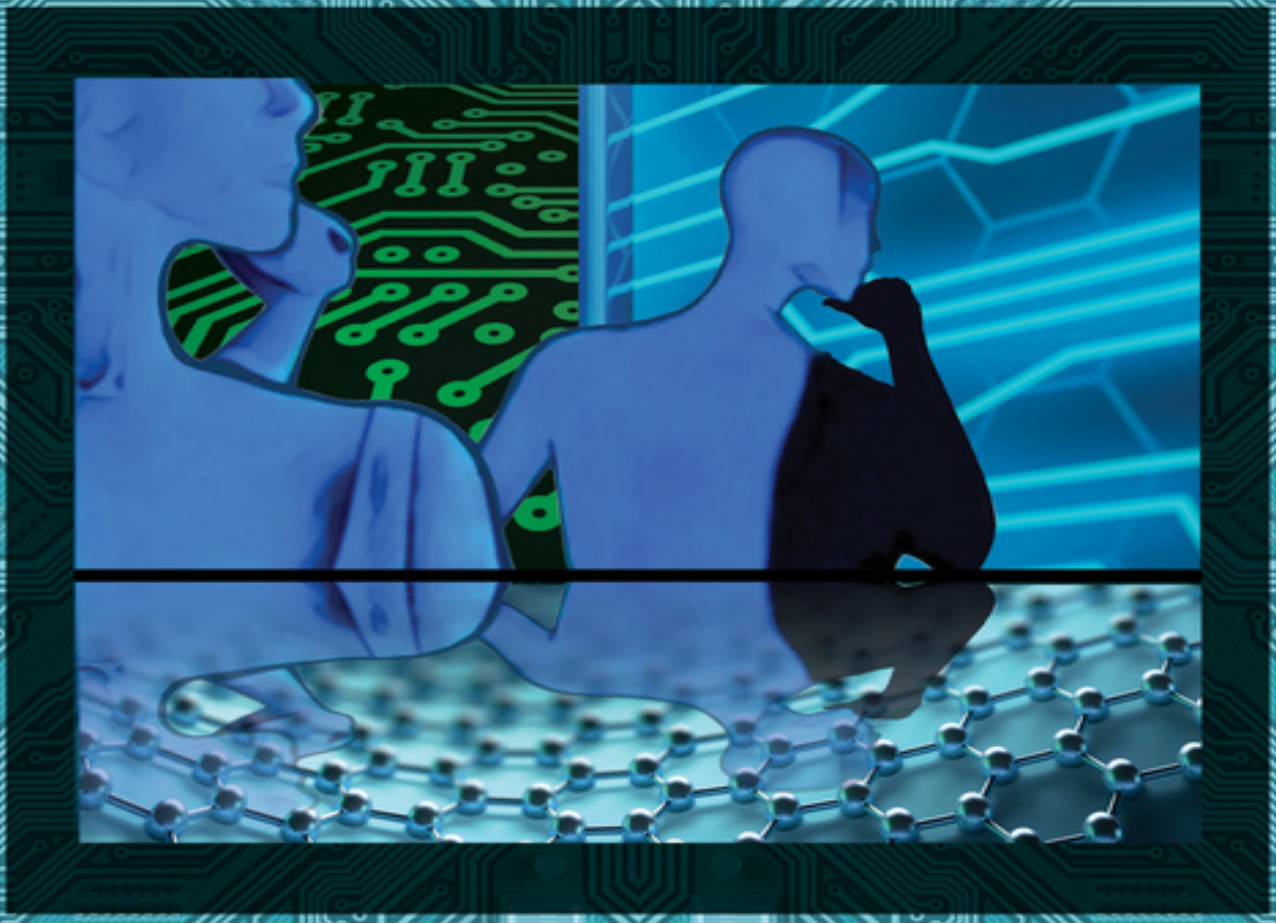
SEMICONDUCTOR BASICS

A qualitative, non-mathematical explanation of how semiconductors work and how they are used



George Domingo

WILEY

Table of Contents

List of Tables

List of Illustrations

Chapter 3

Chapter 6

Chapter 7

Chapter 9

Chapter 11

Semiconductor Basics

A Qualitative, Non-mathematical Explanation of How Semiconductors Work and How They Are Used

George Domingo
Berkeley
CA, USA

WILEY

This edition first published 2020

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Domingo, George, 1937– author.
Title: Semiconductor basics : a qualitative, non-mathematical explanation of how semiconductors work and how they are used / George Domingo, Berkeley, CA, US.
Description: First edition. | Hoboken, NJ, USA : Wiley, [2020] | Includes bibliographical references and index.
Identifiers: LCCN 2020015406 (print) | LCCN 2020015407 (ebook) | ISBN 9781119702306 (cloth) | ISBN 9781119597117 (adobe pdf) | ISBN 9781119597131 (epub)
Subjects: LCSH: Semiconductors. | Solid state electronics. | Electronic apparatus and appliances.
Classification: LCC TK7871.85 .D654 2020 (print) | LCC TK7871.85 (ebook) | DDC 621.3815/2–dc23
LC record available at https://lccn.loc.gov/2020015406
LC ebook record available at https://lccn.loc.gov/2020015407

Cover Design: Wiley
Cover Images: Computer motherboar © Bet_Noire/Getty Images, Two People - Drawn by Pey Llussà, Circuit Board © filo/Getty Images, Central processor unit © Artem_Egorov/Getty Images, Graphene sheet © KTSDESIGN/SCIENCE PHOTO LIBRARY/Getty Images

To my family for their love and support

Acknowledgements

I would like to recognize all those scientists, engineers, professors, authors, teachers, and also students who in the past 130 years with their research, experiments, theories, analysis, publications, and textbook have been able to explain beautifully how matter behaves, how to use it, and how to explain it to the next generation of scientists. In the process they created an electronic revolution. As someone already said, we are sailing on the shoulders of all those thousands of geniuses that preceded us.

I want to acknowledge the efforts and support of the Wiley editors who made the text more readable and clearer.

I want to thank my family who have been always helpful, encouraging, and patient with me and my project.

Finally, I would like to mention Dr. Gerry Hanh, who upon reading the first chapters that I wrote as a hobby, insisted I send them to Wiley, resulting in the book you have now in your hands.

Introduction

A couple of years ago, I was asked to give a talk to the Rotary club in Terrassa, a city about 20 miles northwest of Barcelona. They asked me to divide the talk into three parts: 15 minutes biographical, 15 minutes on my technical work, and 15 minutes about NASA. The first and last 15 minutes of the talk went well, but the technical explanation about how infrared detectors work was disappointing. Yes, they did understand the uses and applications of infrared detectors, my technical work with the astronomical observatories of NASA, but the explanation on how infrared detectors work was not as clear and instructive as I had hoped. I was then and I am now convinced that any educated person can understand how semiconductor devices work. This talk two years ago was my motivation to start writing this book.

Semiconductors are the basis for almost all modern electronic devices. For many people semiconductors are a mysterious material that somehow has taken over modern electronics. In the same way that we understand the concept of god and creatures, but semi-gods are confusing, most of us have an idea of what a conductor (electricity flows through it) and an insulator (it doesn't) are, but what the heck is a semiconductor? Furthermore, the prevalent material used for fabricating semiconductors is pure silicon, the second most common element found on earth (28%) after oxygen (47%). Why not use aluminum, the next common element (8%), or strontium or some other exotic and more classy material? Why do we use semiconductors instead of conductors? Don't we want electrons to move freely through our devices?

I attempt here to explain why and how semiconductors work in a form that any educated person can understand. Every chapter explains the subject in a qualitative way, with drawings, photographs, and figures, and very simple relationships (I hate to use the word "equations" here). At the end of most chapters I add

appendices where I include some mathematical formulation with the relevant equations for those who would enjoy looking a little deeper into the subject. I do not try to proof these equations; my purpose is not formally present to you the physics of semiconductors (there are many excellent books that do that) but just to show what the results are. In this book you have to take these results on faith.

Unless you are very allergic to math, do not skip the appendices. There are interesting concepts that amplify the understanding of how semiconductors work. Don't worry; there are no problems at the end of each chapter and no suggested quizzes.

First, I explain what a semiconductor is, the different types we use, and how they are different from conductors and insulators. Next, I explain the key devices that can be constructed using the semiconductor materials: diodes, passive element, and transistors. I talk about the integrated circuits, how we build them and the larger electronic components, and finally what can we expect in the future (Chapter 15). I interrupt the "theoretical" flow with chapters devoted to applications (Chapters 4, 7, and 9) that can be understood with just the concepts I have covered in the previous chapters.

This book can be read in different ways. If you are interested in understanding how transistors work, then you should read, in succession, Chapters 1, 2, 3, 5, and 8. You cannot skip any of them or read them in a different order unless, of course, you are already familiar with some of the previous topics. Chapter 6 explains the basic electrical components (resistors, capacitors, and inductors) which we need to understand how we build useful semiconductor circuits. I try to discuss some of the semiconductor applications as soon as I cover the relevant theory. Just by understanding the concept of energy levels and energy bands (Chapters 2 and 3) you can grasp how infrared detectors work (Chapter 4). You do not even need to know what resistors or capacitors are. Similarly, after Chapter 6 you can understand the different circuit applications we can fabricate with diodes (Chapter 7) and after Chapter 8 I explain how we use transistors (Chapter 9). The next chapters, integrated

circuit fabrication (Chapter 10) and logic circuits (Chapter 11), can be read separately at any time, although knowing how a transistor works will help. Yes, I talk about semiconductor devices, but you do not need to know the physics of how they work to understand these two chapters.

Next, in Chapter 12, I explain large semiconductor components like multiplexers and memories. In Chapter 13, optoelectronics, I cover lasers and LEDs, and I end the book by discussing some simple concepts related to computer architectures and liquid crystal displays in Chapter 14. I speculate about the future in Chapter 15.

The objective of this book is to explain to the layman how semiconductors work and how they are used. I hope I have succeeded. If you have any comments or if you find errors, inconsistencies, unclear, incomplete or confusing figures or explanations, please let me know by emailing me at semiconductorbasics@hotmail.com.

I had a professor who taught me relativistic quantum mechanics, one of the hardest courses I ever took, who used to say, "Everything mathematical is trivial." I agree with him, not 100%, but partially so. The truth is that we scare the hell out of young students by making them believe that mathematics, physics, and engineering are difficult. Wrong! Math is not difficult and physics, and all its derivatives, is fascinating.

March 2019

So please enjoy.
George Domingo

1
The Bohr Atom

OBJECTIVES OF THIS CHAPTER

To understand how semiconductors work and how they are used, we need to be familiar with the concept of allowed energy levels first proposed by Niels Bohr. How Bohr came up with the idea of a planetary model of the atom is very interesting. Science is a continuum: one observation leads to a hypothesis that leads to a theory that leads to more observations, and so on. Bohr did not come up with his model of the atom out of the blue – no apple fell on his head. A lot of observations and theories going back to the 1700s were proposed before he put them together in a now well-known theory.

In this chapter, I discuss the experimental evidence and the scientific observations that led to Bohr's planetary model of the atom and the discrete energy levels it postulated. This brief journey will help us understand the significance of the Bohr atom that explains the strange behavior of light spectra.

1.1 Sinusoidal Waves

Before I start, I want to clarify the terms used to define a wave, which I use in the next few sections, and the relations between these terms (see Figure 1.1). There are four variables that we use to define any sinusoidal wave:

The *amplitude*, *A*: How high each peak of the wave is in relation to the middle, its zero value.

The *frequency*, *f*: The number of ups and downs in the wave in a given time. The units are Hertz or number of ups and downs per second: a number/s.

The *wavelength*, λ: The distance between two peaks, in meters (m), centimeters (cm), or micrometers (μm).

The *wave number*, ν (the Greek letter nu, not the letter *v*): The reciprocal of the wavelength. Some properties of waves are better expressed by the reciprocal of the wavelength. The units are therefore 1/m or m^{-1}, or cm^{-1}, or μm^{-1}.

The last three variables are related by the velocity of the wave. *Velocity* is the distance that a moving object covers during a fixed amount of time, so the velocity *v* (this is now the letter *v*) has units of meters per second (m/s). The key relationships are

$$f = \frac{v}{\lambda} \quad \text{or} \quad \lambda = \frac{v}{f} \tag{1.1}$$

and the wave number – the reciprocal of the wavelength – is

$$\nu = \frac{1}{\lambda} = \frac{f}{v} \tag{1.2}$$

For example, suppose that Figure 1.1 represents a sound wave. The velocity of sound in air is 343 m s^{-1}. Take a look at Figure 1.1:

The figure shows 5 cycles in 0.001 seconds, which means the frequency is 5000 cycles per second or f = 5000 Hz, (where Hz, Hertz is the unit for frequency) which happens to be the middle of our hearing range.

The wavelength is the velocity divided by the frequency, or λ = 343 (m s^{-1})/5000 (1 s^{-1}) = 0.069 m or 6.9 cm. Notice that the seconds cancel out, and therefore the units are in meters or centimeters.

The wave number is ν = 1/0.069 m = 14.5 m^{-1}.

As much as possible, I use the metric system of units (MKS, meter, kilogram, second). I have always found it very annoying when books keep changing the unit system. When necessary, I will give you the equivalents.

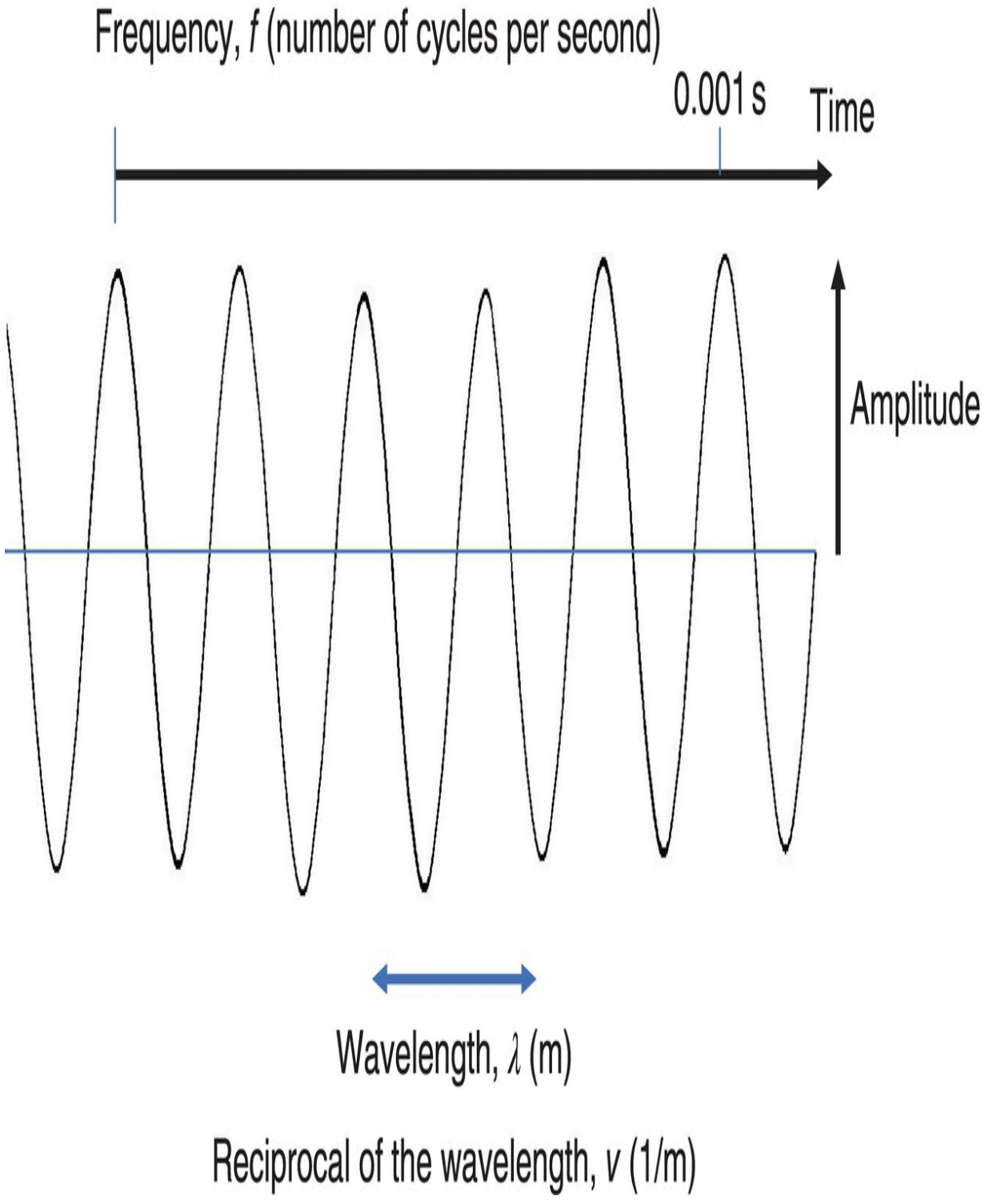


Figure 1.1 A sinusoidal wave is described in several ways: frequency, wavelength, and reciprocal of the wavelength plus its amplitude.

Now we are ready to dive into the pre-history of the Bohr atom and understand how Dr. Bohr came up with his famous model.

1.2 The Case of the Missing Lines

To explain how semiconductors work, we start with the Bohr atom. Most readers are familiar with Bohr's planetary model of the atom. How Bohr came up with this model is a very interesting scientific historical path involving many famous scientists of the nineteenth and early twentieth centuries. Science goes one step at a time.

William Wollaston (1766–1826), shown on the left in Figure 1.2, was an English chemist who discovered a couple of atomic elements, including palladium and rhodium. Very early in the 1800s, he built the first spectrometer. Wollaston focused the light from the sun through a prism and, to his surprise, found black lines partitioning the spectra (Figure 1.3). What the heck was going on?

Engraved by W. Holl.

W. H. WOLLASTON.

From the original Picture by J. Jackson
in the possession of the Royal Society.

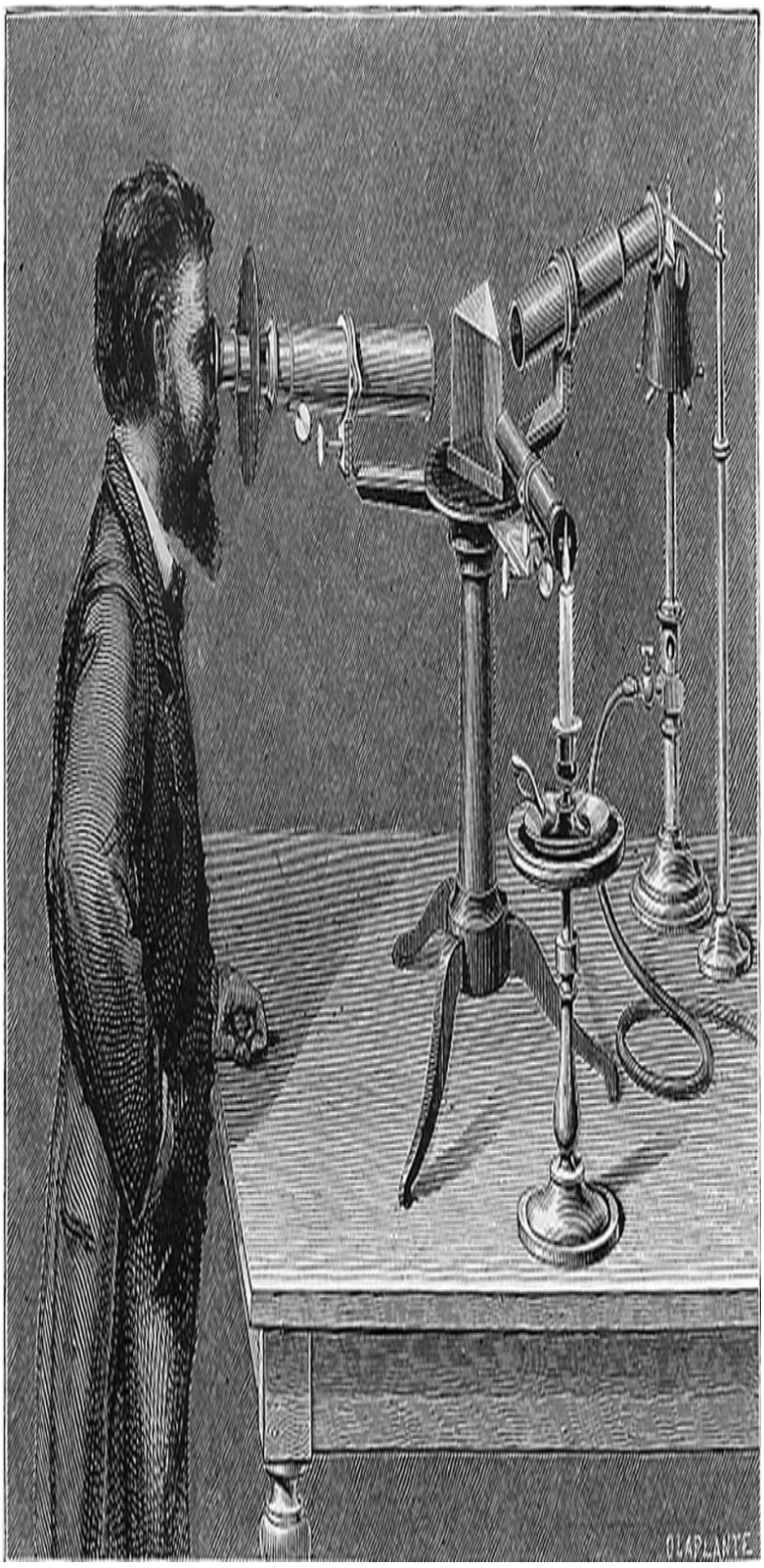


Figure 1.2 William Wollaston (left) looked at the sun's light through a prism and was the first to observe the missing lines.

Source: https://library.si.edu/image-gallery/73731. Joseph von Fraunhofer (right) studied the missing lines with his spectrometer and named them A–K. where Hz, Hertz is the unit for frequency. https://www.kruess.com/en/campus/spectroscopy/history-of-spectroscopy/

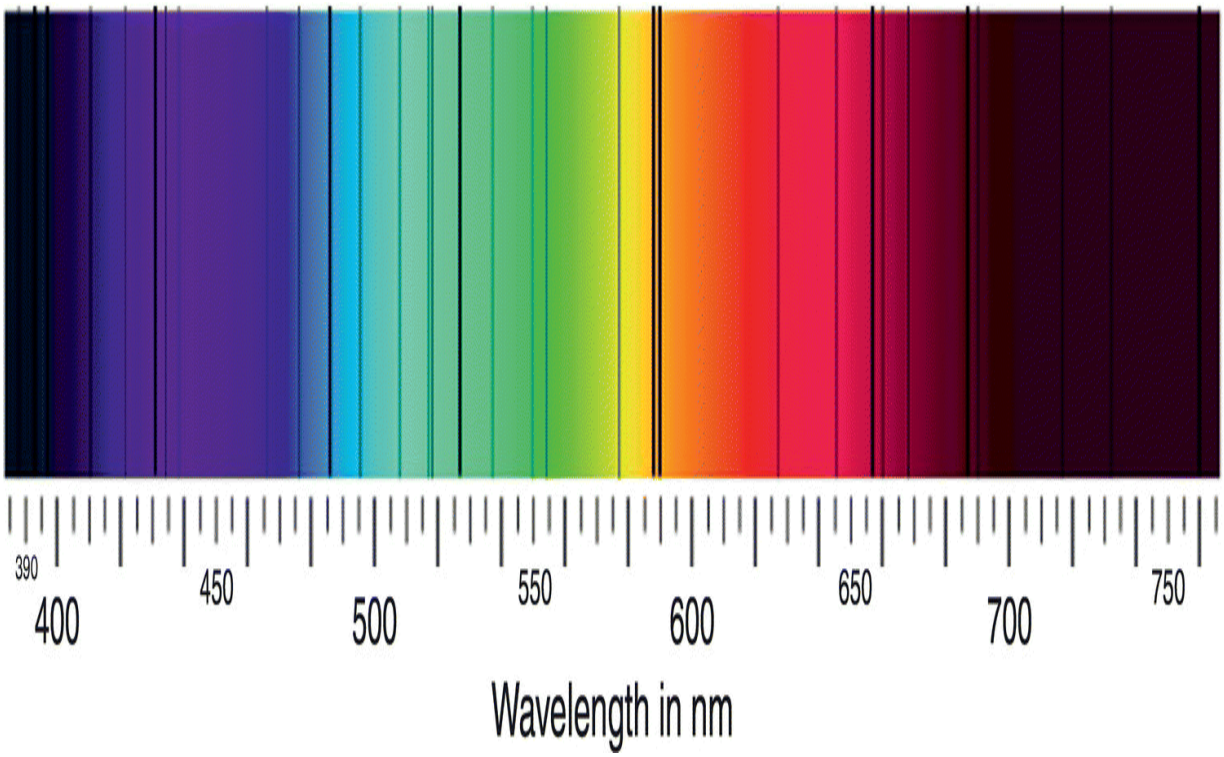


Figure 1.3 The sun's spectrum through a prism shows dark lines: wavelengths of light that seem to have disappeared.

Source: https://www.kruess.com/en/campus/spectroscopy/history-of-spectroscopy/.

Suppose you are roasting a chicken and carefully watching the dial of a digital thermometer inserted in the chicken's breast as the temperature increases from room temperature, 24 °C (degrees Celsius), to 80 °C, the recommended internal temperature of a well-cooked chicken breast. As the temperature increases, suddenly the thermometer jumps from 39 °C to 41 °C, then from 56 °C to 58 °C, and finally from 66 °C to 68 °C. You wonder what is wrong with the thermometer: why don't the temperatures 40, 57, and 67 °C show

up on the dial? They don't seem to exist. You buy a new thermometer, just to be sure, and find that exactly the same temperature values are missing. A third thermometer gives the same results. You place the same thermometers in soup, and the thermometers are well behaved, showing in succession the values 39, 40, and 41 °C. So, the thermometers work. The missing temperatures are no coincidence. There is something in that chicken that makes the temperature jump from one value to another without passing through the one in the middle.

Well, that was probably Wollaston's initial reaction. What separated the colors? Was the instrument lens dirty? He even considered the possibility that there were natural boundaries between certain colors. But why didn't these black boundaries appear when he pointed the spectrometer at a white light?

German physicist Joseph von Fraunhofer (1787–1826), on the right in Figure 1.2, studied these dark lines of the sun's spectrum in much more detail and actually named the missing lines with the letters A–K (not too imaginative; ancient astronomers would have found much more attention-grabbing names).

1.3 The Strange Behavior of Spectra from Gases and Metals

The next step was to take a look at the emission of light from several metals, gases, and stars. The observers soon found that each element has its own unique set of lines. Take a look at Figure 1.4. White light through a prism generates a continuous spectrum (top diagram); no colors are missing. If we heat a gas until it glows and pass the light through the same prism (middle diagram), only certain lines are projected onto the screen: the rest of the spectrum has disappeared. But if we do it differently – that is, if we have the cold gas between the white light and the prism (bottom diagram) – then the full spectrum appears, except for *exactly* the same lines that were visible in the previous spectral measurement. The

superposition of the middle and lower spectra is equal to the spectrum of the white light at the top. The gas, when cold, absorbs the same specific light waves as the ones it emits when it glows. What a coincidence! And this happens with *all* gases and materials. The lines are at different frequencies, but all of them have lines.

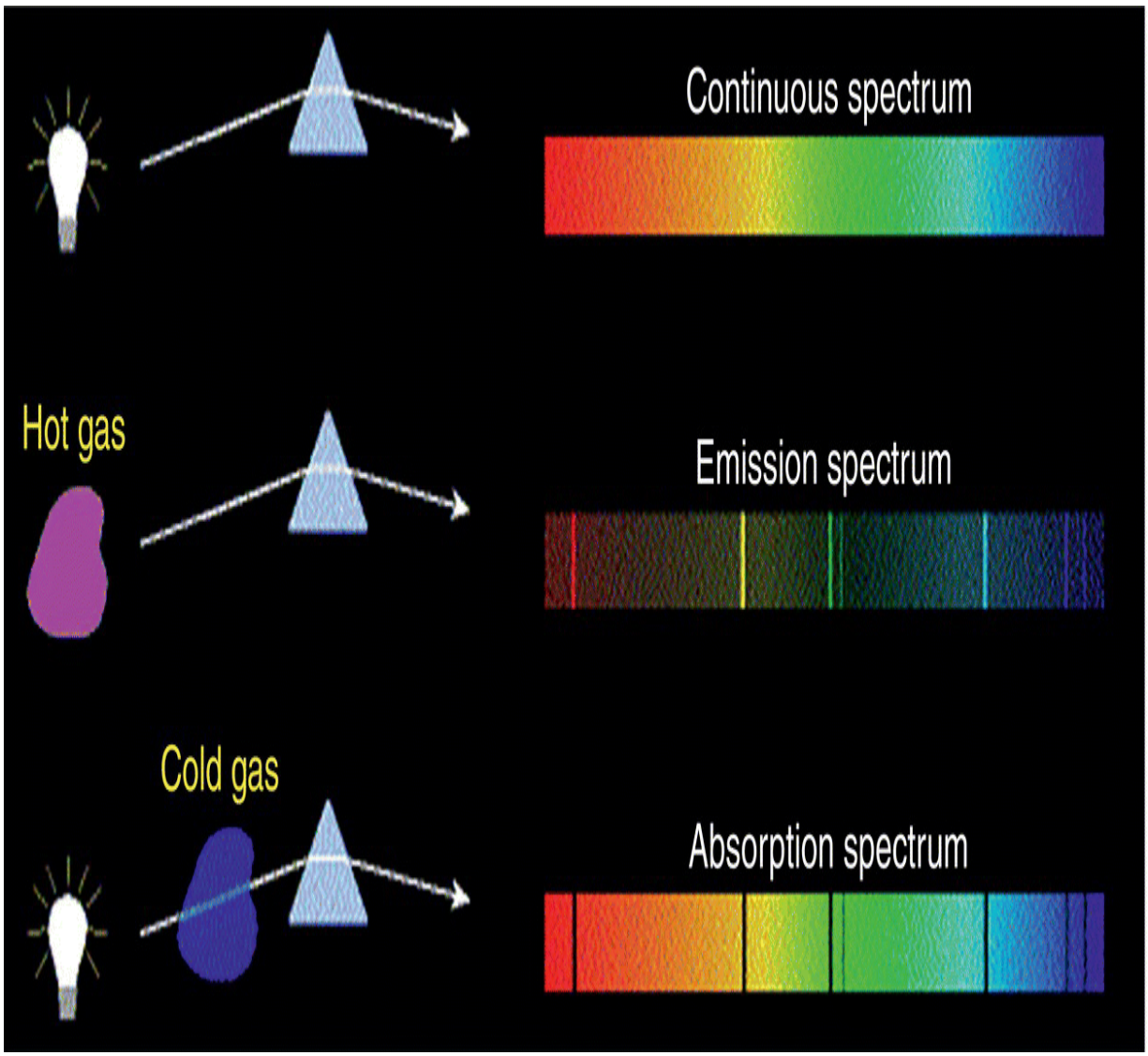
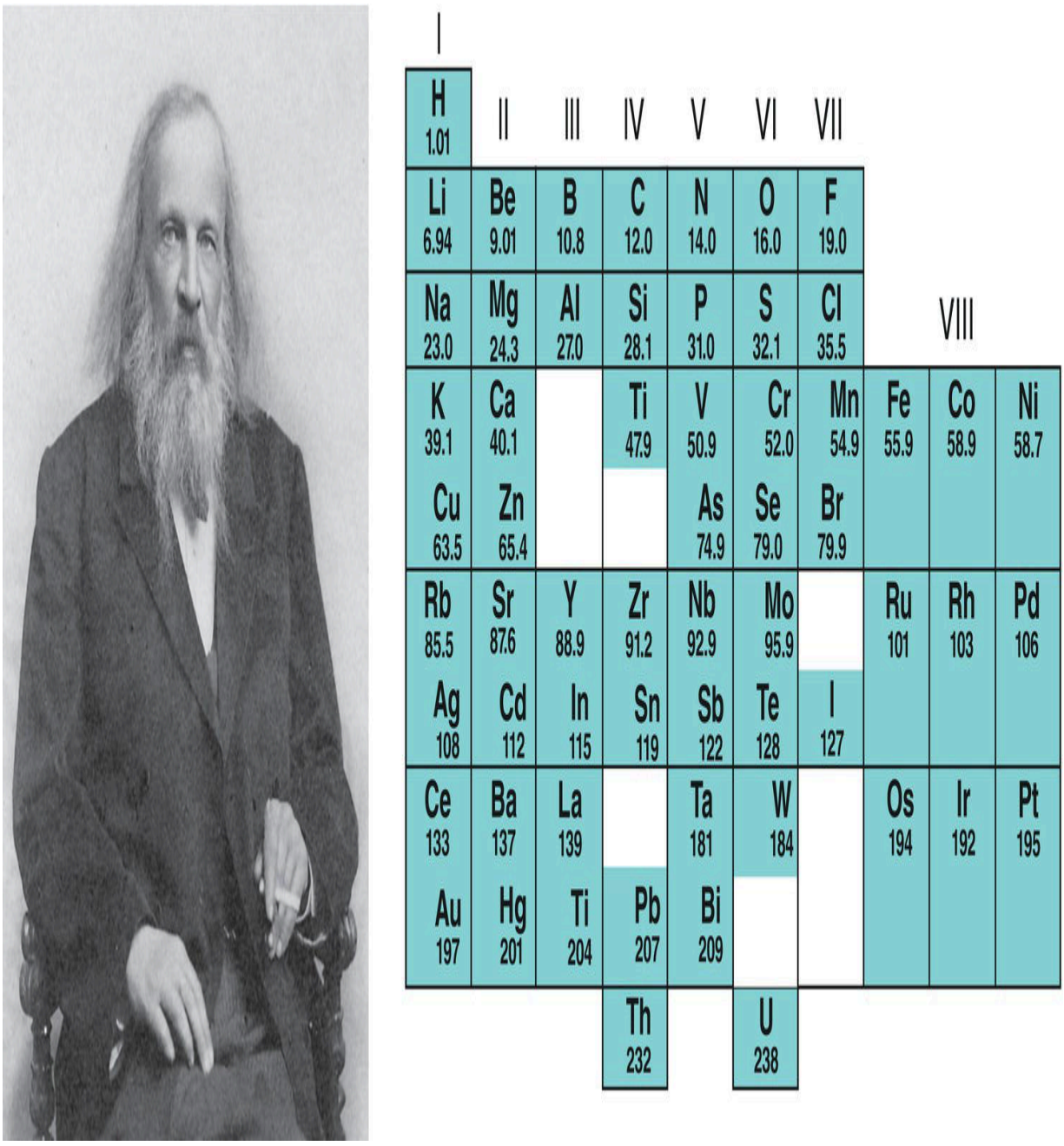


Figure 1.4 The spectrum from any gas shows similar but different missing lines (middle image), but when the same gas is hot and emits light, only the lines that were black before are now visible (lower image).

Source: https://quizlet.com/102018176/astronomy-4-spectroscopy-flash-cards/.

1.4 The Classifications of Basic Elements

In 1766, an English aristocrat named Henry Cavendish (1731–1810) was the first to recognize hydrogen as an element, that is, a unique substance and an integral part of the water molecule. By the time Dmitri Mendeleev (1834–1907) published his now ubiquitous periodic table of the elements in 1869, 100 years later, it was already known that hydrogen was the lightest of all the known elements (at that time, 60 elements were known). Mendeleev (Figure 1.5) reordered the known elements by their relative atomic weight, with hydrogen as 1.



I	II	III	IV	V	VI	VII	VIII		
H 1.01									
Li 6.94	Be 9.01	B 10.8	C 12.0	N 14.0	O 16.0	F 19.0			
Na 23.0	Mg 24.3	Al 27.0	Si 28.1	P 31.0	S 32.1	Cl 35.5			
K 39.1	Ca 40.1		Ti 47.9	V 50.9	Cr 52.0	Mn 54.9	Fe 55.9	Co 58.9	Ni 58.7
Cu 63.5	Zn 65.4			As 74.9	Se 79.0	Br 79.9			
Rb 85.5	Sr 87.6	Y 88.9	Zr 91.2	Nb 92.9	Mo 95.9		Ru 101	Rh 103	Pd 106
Ag 108	Cd 112	In 115	Sn 119	Sb 122	Te 128	I 127			
Ce 133	Ba 137	La 139		Ta 181	W 184		Os 194	Ir 192	Pt 195
Au 197	Hg 201	Ti 204	Pb 207	Bi 209					
			Th 232		U 238				

Figure 1.5 Dmitri Mendeleev and the periodic table with the elements known in his time and the empty slots for elements still to be discovered.

Source: Wikipedia, https://en.wikipedia.org/wiki/Dmitri_Mendeleev#/media/File:Dmitri_Mendeleev_1890s.jpg.

1.5 The Hydrogen Spectrum Lines

Johann Balmer (1825–1898), a Swiss mathematician, found empirically in 1895 that the separation of the optical lines generated by hydrogen gas can be expressed by a formula using just a constant, *C*, and integer numbers. He expressed his observation with Eq. (1.3)

$$\lambda = \frac{Cm^2}{m^2 - n^2} \tag{1.3}$$

here λ is the wavelength of the missing line, C is a heuristically obtained constant ($C = 3.64 \times 10^{-9}$ m), $n = 2$, and m is an integer greater than 2 (i.e. 3, 4, 5, and so on). When you put any of these integer numbers in Eq. (1.3), you get the wavelength of all the lines in the hydrogen spectrum.

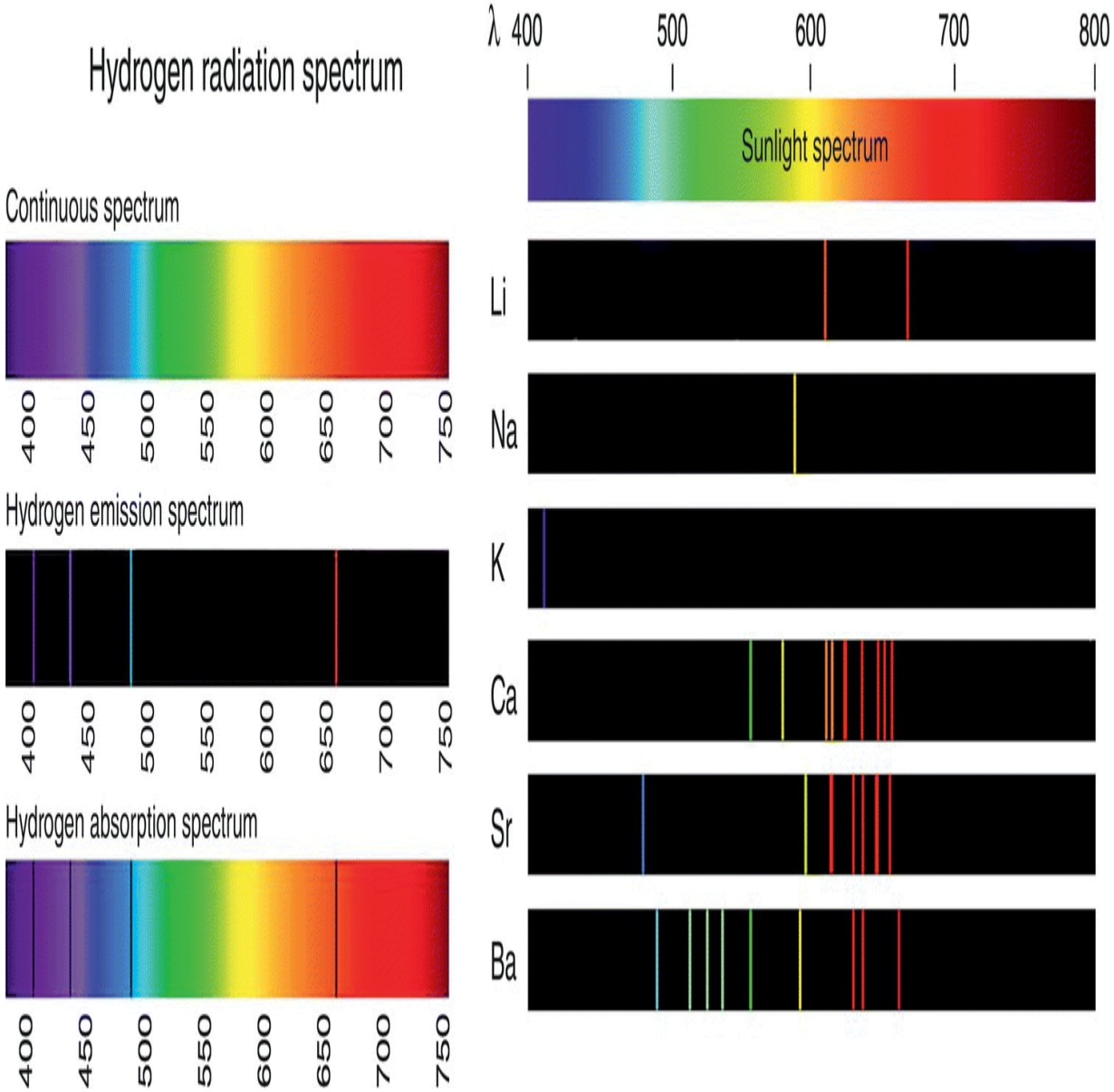


Figure 1.6 The spectrum of the hydrogen atom on the left shows the absorption lines (below) and the emission lines (middle). On the right are the emission lines of several other materials.

Source: https://www.shutterstock.com/image-vector/spectrum-spectral-line-example-hydrogen-emission-1288942888?src=iUiOwiDEznOcV6XzswXhMA-1-0 (left); https://www.shutterstock.com/image-vector/line-spectra-elements-339037577?src=I6tWF1qlh6XcWayXsZl-Gw-3-16 (right).

Figure 1.6 shows the hydrogen spectrum on the left, with its characteristic emission and absorption lines. These are the lines that

Balmer used to develop Eq. (1.3) to calculate the missing hydrogen's wavelengths. All the elements have similar absorption and emission lines at different wavelengths, and I show a few on the right in Figure 1.6.

Just three years later, Johannes Rydberg (1854–1919) found that the Balmer equation was one specific case of a more general formula, Eq. (1.4):

$$\nu = \frac{1}{\lambda} = R\left[\frac{1}{n^2} - \frac{1}{m^2}\right] \tag{1.4}$$

The reciprocal of the wavelength is now given by a constant R and the same integer numbers, except that now n is allowed to have different integer numbers: 2, 3, 4, and so on. R is also a heuristically derived constant ($R = 1.1 \times 10^7\ \mathrm{m}^{-1}$), called the *Rydberg constant*. Both Balmer and Rydberg (Figure 1.7) were able to quantify the entire spectrum of the hydrogen atom using the relationship in Eq. (1.4). It is interesting that Niels Bohr, whom I'll talk more about in Section 1.8, was able to calculate the Rydberg number using fundamental physical values, such as the mass of the electron, the electronic charge, the permittivity of free space, Planck's constant, and the speed of light (see Appendix 1.3). This behavior screams for an explanation.

Figure 1.7 Johann Balmer (left) found a mathematical relation for hydrogen's spectral lines, and Johannes Rydberg (right) came up with a more general equation applicable to all gases and materials.

Source Wikipedia, https://en.wikipedia.org/wiki/Johann_Jakob_Balmer#/media/File:Balmer.jpeg (left); Wikipedia, https://en.wikipedia.org/wiki/Johannes_Rydberg#/media/File:Rydberg,_Janne_(foto_Per_Bagge;_AFs_Arkiv).jpg (right).

1.6 Light is a Particle

Albert Einstein (1879–1955, Figure 1.8) published a paper in 1905 on the theory of the photoelectric effect. When light strikes a metal surface, it frees an electron if its energy is higher than a given threshold value. Any remaining energy is used to kick the electron off the surface. In his paper, Einstein proposed the concept that light has a dual personality; it behaves like a wave or like a particle, and the particle has an energy associated with the wavelength of that light.

He called this particle a "light quantum." (In 1926, a French physicist named Frithiof Wolfers [1891–1971] renamed the light quantum a *photon*. It is interesting that Einstein received the Nobel Prize in 1921 for the discovery of the photon, not for his much more famous work on relativity.) This light particle, the photon, has an energy that depends on the frequency of the light. The energy associated with this light is given by the formula

$$E = hf = \frac{hc}{\lambda} \tag{1.5}$$

where h is Planck's constant ($h = 6.63\,10^{-34}\ \mathrm{m^2\,kg\,s^{-1}}$), c is the speed of light ($c = 3 \times 10^8\ \mathrm{m\,s^{-1}}$), and λ is the wavelength (m). The meter in the numerator cancels the one in the denominator, resulting in the energy given in Joules ($= \mathrm{kg\,m^2\,s^{-2}}$).

Figure 1.8 Around 1905, Albert Einstein came up with the concept that light behaves as both a wave and a particle.

Source: Wikipedia, https://en.wikipedia.org/wiki/Albert_Einstein#/media/File:Einstein_patentoffice.jpg.

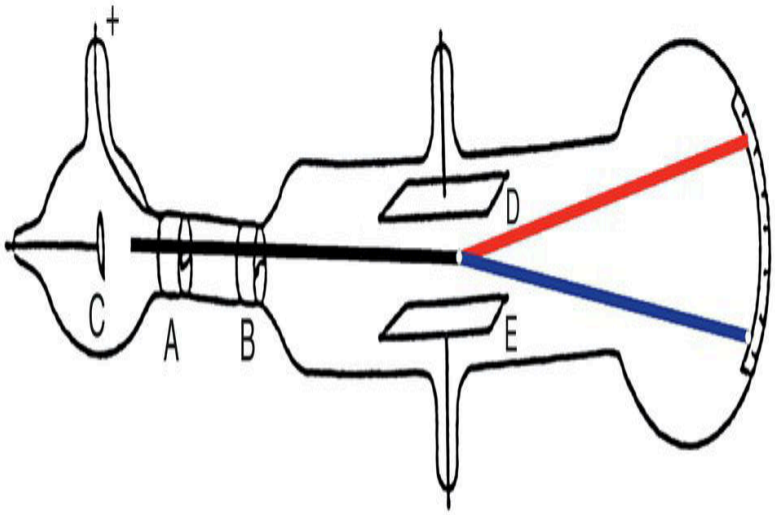
1.7 The Atom's Structure

While all of these light experiments and relationships were being observed in the late nineteenth century, other scientists were playing with cathode-ray tubes, the precursors of old television sets and oscilloscopes, trying to understand the nature of the atom. The cathode-ray tube consists of an evacuated tube with two contacts, one at each end: the *cathode* and the *anode*. When a voltage is applied across the tube, current flows from the cathode to the anode, and the tube glows. The scientists explained this phenomenon by saying that electrons going through an evacuated tube containing very few atoms are able to attain sufficient velocity (and therefore kinetic energy) to hit the atoms and make them glow. They were called *cathode rays*.

Nobel Prize winning British physicist Joseph John Thomson (1856–1940, Figure 1.9) studied cathode rays and postulated in 1897 that they consisted of extremely small negatively charged particles, which he initially called "corpuscles." (As happened with the term *photon*, George Stoney (1826–1911) later renamed corpuscles as *electrons*.) By studying how these particles moved through the gas and how they could be deflected by magnets, Thomson concluded that the "corpuscles" were (i) negatively charged particles and (ii) much smaller than the atoms themselves – at least 1000 times smaller. To account for electrically neutral atoms, he proposed that there is a core of positive charges with a large mass surrounded by an amorphous cloud of negatively charged electrons.

Ernest Rutherford (1871–1937, Figure 1.10), also a Nobel Prize winner, worked with radioactivity. In 1911, he bombarded very thin gold foil with alpha particles and looked at the scattered reflections

as the radiation went through the foil. Most of the radiation went through the foil undeflected. Only a few alpha particles were reflected back and, from the angle of the reflected radiation, he concluded that the atom must have a very small, concentrated, positively charged core to compensate for the negatively charged electrons. Because the large majority of alpha particles passed through the foil without any directional change, he concluded that the majority of the space in an atom is empty, and the electrons are orbiting the nucleus instead of just being a scrambled negatively charged cloud as Thomson had suggested.



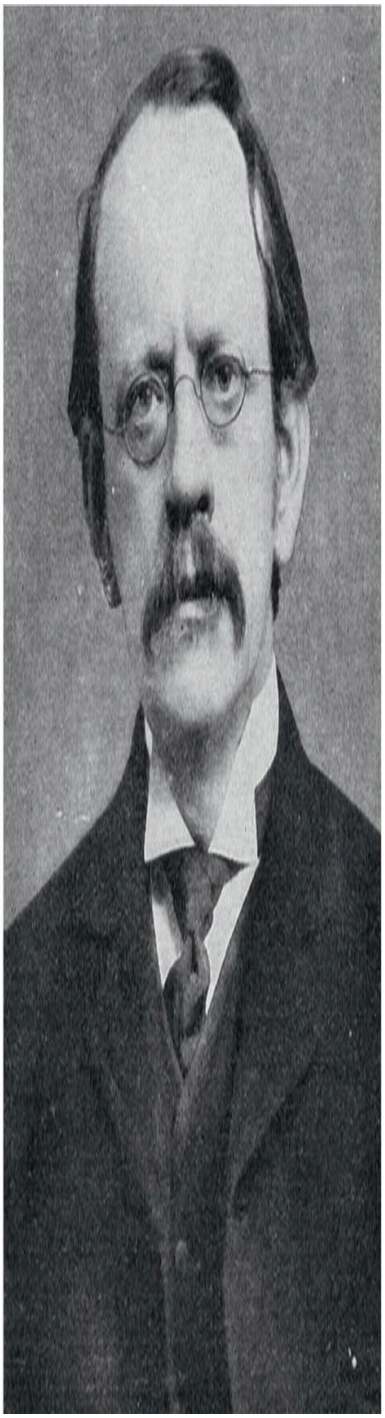
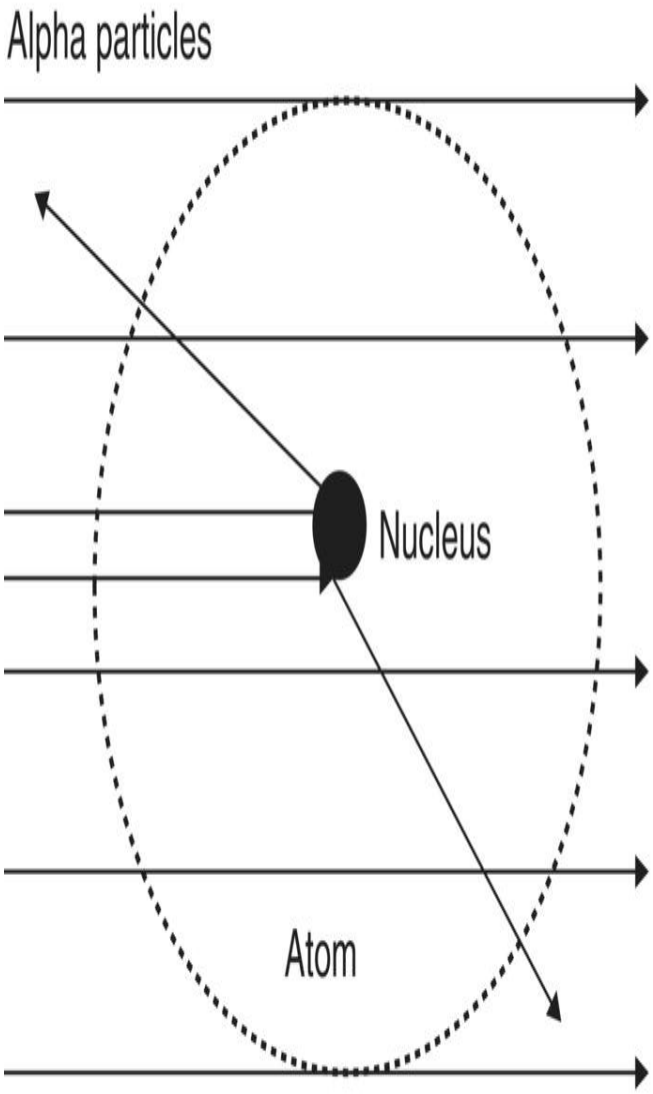


Figure 1.9 Joseph John Thomson and his cathode ray tube.

Source: Wikipedia, https://en.wikipedia.org/wiki/J._J._Thomson#/media/File:JJ_Thomson_Cathode_Ray_2.png (left); Wikipedia, https://en.wikipedia.org/wiki/J._J._Thomson#/media/File:J.J_Thomson.jpg (right).


Alpha particles
Nucleus
Atom

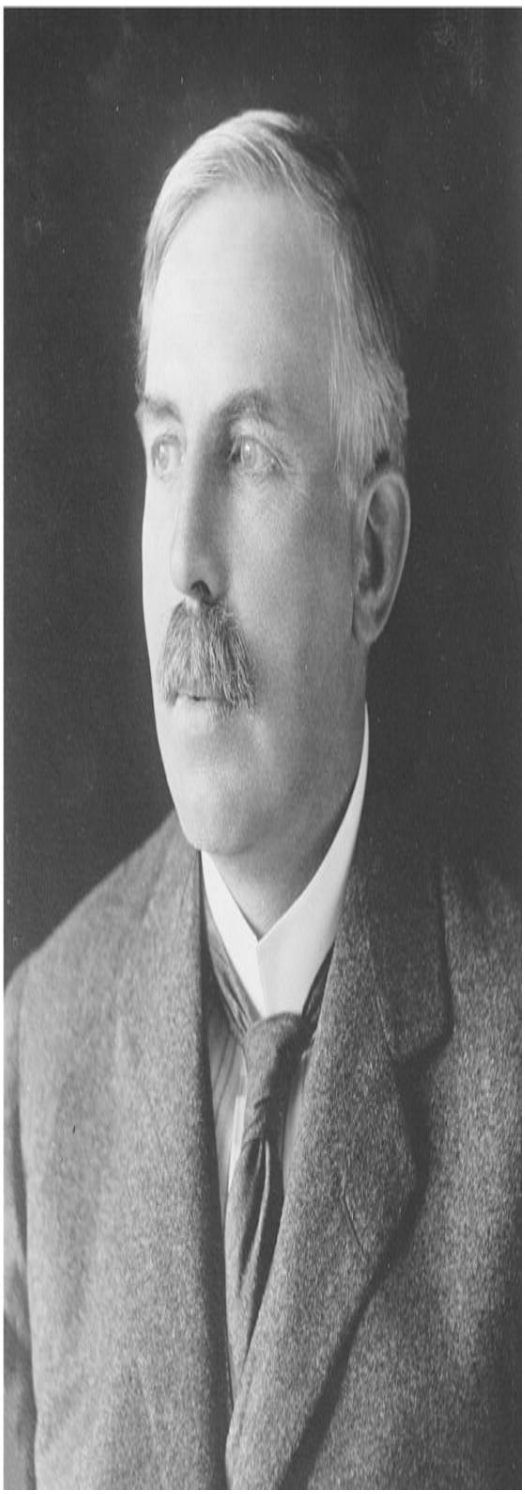


Figure 1.10 Ernest Rutherford, with his experiment that bombarded alpha particles with radiation, concluded that the nucleus is extremely small and is concentrated at the center of the atom.

Source: Wikipedia, https://upload.wikimedia.org/wikipedia/commons/6/6e/Ernest_Rutherford_LOC.jpg.

Robert Millikan (1868–1953, Figure 1.11) was able to measure the electrical charge of an electron with an interesting oil drop experiment in 1909. He suspended a very small charged oil drop between two metal plates – one positive and the other negative – creating an electric field between the plates. He dropped tiny oil droplets into a vacuum chamber, and with X-rays, he negatively charged some of the oil drops. By changing the electric field between the two plates, he could control the speed of the oil drops, slowing them down, stopping them, or even moving them upward. By knowing the density of the oil drop, the size of the drop, its volume and mass, and the electric field that compensated for the effect of gravity, he was able to come up with the value of the charge of a single electron: 1.592×10^{-19} coulombs (he was off by less than 1% of the now-established number – not bad at all).

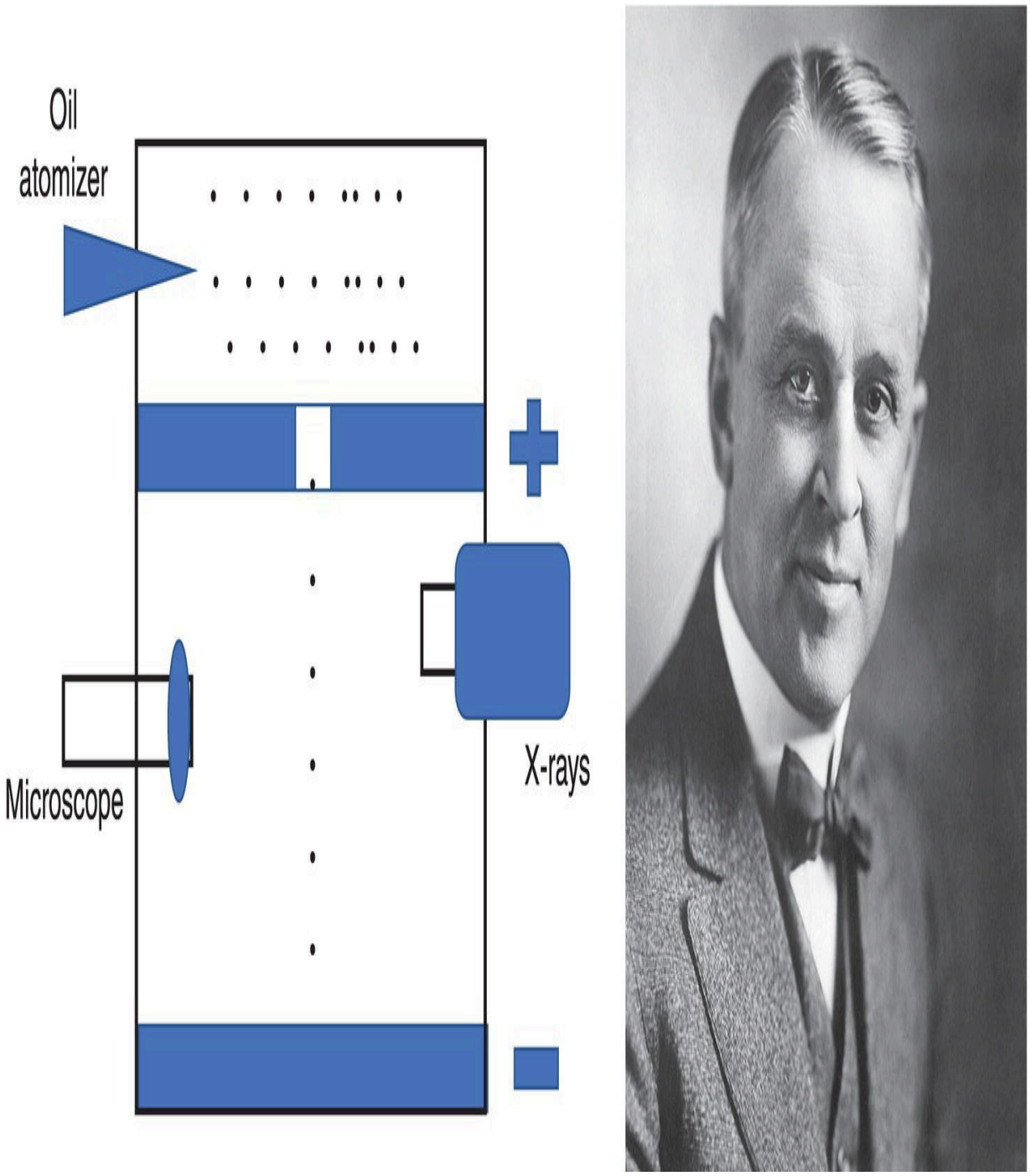


Figure 1.11 Robert Millikan, with his oil-drop experiment, measured the electrical charge of an electron.

Source: https://en.wikipedia.org/wiki/Robert_Andrews_Millikan#/media/File:Millikan.jpg .

1.8 The Bohr Atom

So here we are in 1913 (just a mere 105 years ago at the time of this writing). What did Bohr know? He knew:

That a hydrogen atom is the simplest atom, consisting of just one proton (positively charged) and one electron (negatively charged).

That all of the atom's mass is concentrated at the core: that is, the proton.

That electrons are negative particles somehow orbiting the nucleus.

That the great majority of space in an atom is empty.

That all the other elements can be organized neatly by weight on a periodic table.

That all elements have different emission spectra with specific emission or absorption color lines.

Niels Bohr (1885–1962, Figure 1.12) was able to beautifully explain all of these observations and how the spectral lines are generated. He postulated in 1912 that an atom consists of a core nucleus that has all the mass and is surrounded by electrons, moving like a planetary system in well-defined orbits (Figure 1.13). Electrostatic forces between the proton and the electron (analogous to the gravitational forces in the solar system) keep the electrons circulating without escaping their orbit. Additionally, Bohr postulated that the electrons in orbit do not radiate any energy, so the orbits are stable. The only way to radiate or absorb any energy is for an electron to jump from one orbit to another, and that is precisely what explains the spectra of hydrogen and other elements.

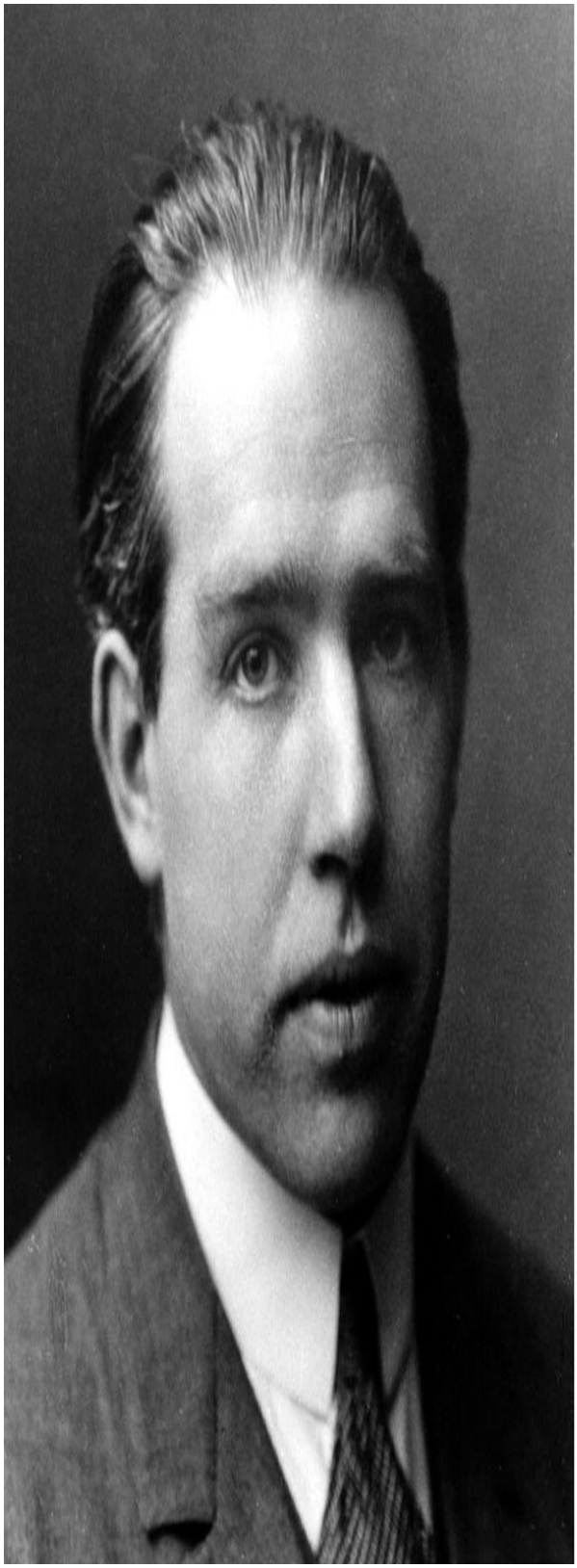


Figure 1.12 Niels Bohr (left) postulated the planetary model of the atom. Wolfgang Pauli (right), using quantum mechanics, proved that no two electrons in a system can have the same quantum numbers.

Source: Wikipedia, https://en.wikipedia.org/wiki/Niels_Bohr#/media/File:Niels_Bohr.jpg (left); Wikipedia, https://en.wikipedia.org/wiki/Wolfgang_Pauli#/media/File:Pauli.jpg (right).

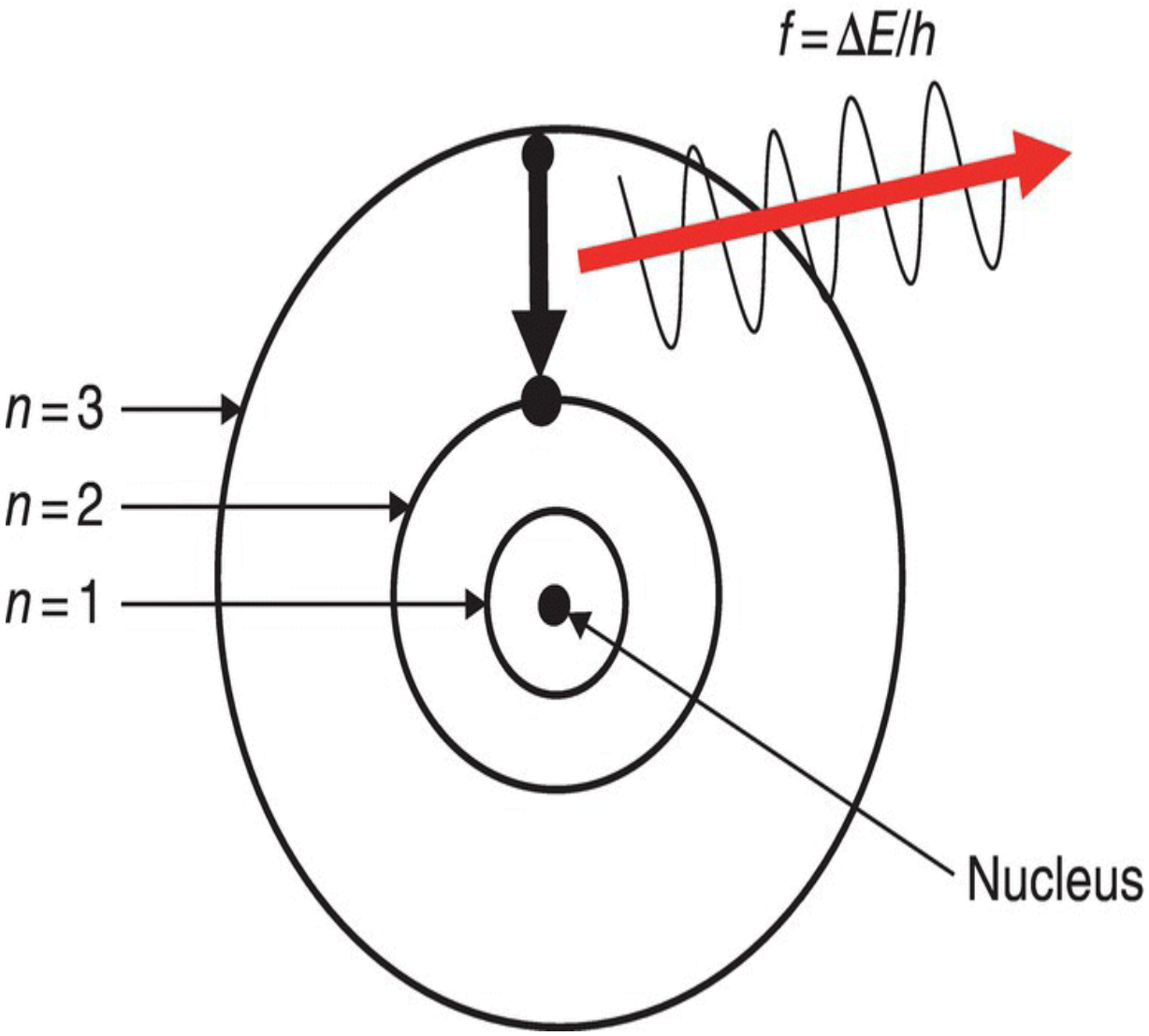


Figure 1.13 The Bohr planetary model of an atom has discrete and stable orbits. An electron falling from level 3 to level 2 transfers its energy to an equivalently energetic photon.

Since electrons are forbidden to have any energy except for the

energy of a specific orbit, they have to jump from one orbit to another, like going up the stairs, one, two, or three steps at a time (not one and a half). When falling from a higher orbit to a lower one, the electron releases a fixed packet of energy in the form of a photon of a very precise frequency (remember that Einstein said light behaves like a particle with an energy related to the wavelength of the light: Eq. 1.5). The transition from orbit 3 to orbit 2, as I show in Figure 1.13, results in the emission of a photon of a very precise frequency, given by the change in energy, ΔE, divided by Planck's constant. Similarly, if an electron in orbit 2 wants to jump to orbit 3, the hydrogen atom has to absorb the energy it needs by absorbing a photon with the same precise energy, or by thermal heating, or by some other means. All other light photons not exactly matched to the difference between the energy levels go through the material unimpaired. The material is therefore transparent for all of the light waves that do not match the exact difference between two energy levels.

In 1924, Austrian Wolfgang Pauli (1900–1958, on the right in Figure 1.12) proposed his exclusion principle, which states that no two electrons (or fermion particles) in a system can have the same quantum numbers. The first atomic level of any element can hold only 2 electrons, the second 8, the third 18, the fourth 32, etc. A simple relation tells you how many electrons can share a given energy orbit: $2n^2$. You may wonder why. If, according to Pauli's exclusion principle, the electrons cannot share the same quantum state, why do we have more than one electron in each orbit? The answer is that each electron is described by four quantum numbers (like the three numbers that describe your first, middle, last names, and your date of birth), but only the first quantum number, n, specifies the energy of the electron and thus explains the behavior of the light spectra. I explain the electron's four quantum numbers in more detail in Appendix 1.1.

Here's an analogy. Suppose I have a theater with 2 seats in the first row, 8 in the second, 18 in the third, 32 in the fourth, etc. The

tickets for the first row cost $20, the second row $50, the third row $75, the fourth $125, and so on. (I know, it is a weird theater, but this is just an analogy.) Spectators are forbidden to sit in someone else's lap or stand in the aisles. If 12 people show up for the performance, they first occupy the 2 seats of the first row, the next 8 patrons occupy the second row, and the last 2 spectators sit somewhere in the third row. Further back in the theater, the rows have more seats, but they are empty. If a patron wants to change rows – from row 3 to row 4, for example – he has to pay the extra $50: the difference in the price of the tickets in the different rows. If he moves the other way, from row 4 to row 3, he is reimbursed the $50. Now, if a wealthy person in row 1 wants to move to row 4, he will be required to pay $105. That is, money must be paid or received to move from one row to another. All these changes assume that the seat someone desires is unoccupied. If the group of spectators is short of money (no energy), they will occupy the seats of the first rows as long as there are seats available. If the group is wealthy (has lots of energy), they can jump from one seat to another as long as they have enough money (energy) to afford the higher prices. The amount they have to pay depends only on the difference in the price of the seats in each row. End of analogy.

At 0 K, absolute temperature (−273 °C), there is no energy whatsoever, so all the electrons occupy the lowest allowed energy levels. At room temperature, 300 °C, there is quite a large amount of thermal energy, and electrons start moving from one level to another, leaving empty seats that can be occupied by other electrons, absorbing or emitting photons as they move.

Figure 1.14 shows the transitions observed in the hydrogen atom. The groups of lines were named later by those who found them.

Have you ever wondered why, when we walk on the second floor, we do not fall through it and land on the first floor? Think about it. The typical size of an atom is 5×10^{-10} m, and the size of a nucleus is about 30 000 times smaller, 1.6×10^{-15} m. All the mass is concentrated in the nucleus. The atoms are, for all practical

purposes, composed of empty space. So why does the empty space of my shoes do not go through the empty space of the tiles on the second floor? It is not due to electrostatic repulsion. Both the soles of my shoes and the tiles are electrically neutral. The reason we do not fall through the floor is the Pauli exclusion principle. The electrons in the sole cannot find a lower energy level on the atoms of the tile. The Pauli exclusion principle not only keeps us safe on the second floor but also explains why material physical objects have any volume at all. It also explains friction. The atoms of the sole locate themselves in a preferential position with the atoms of the floor, and they resist moving. How intense the friction is depends on the crystallographic structure of the two surfaces (Emily Conover, "Giving Friction the Slip", *Science News*, 3 August 2019).

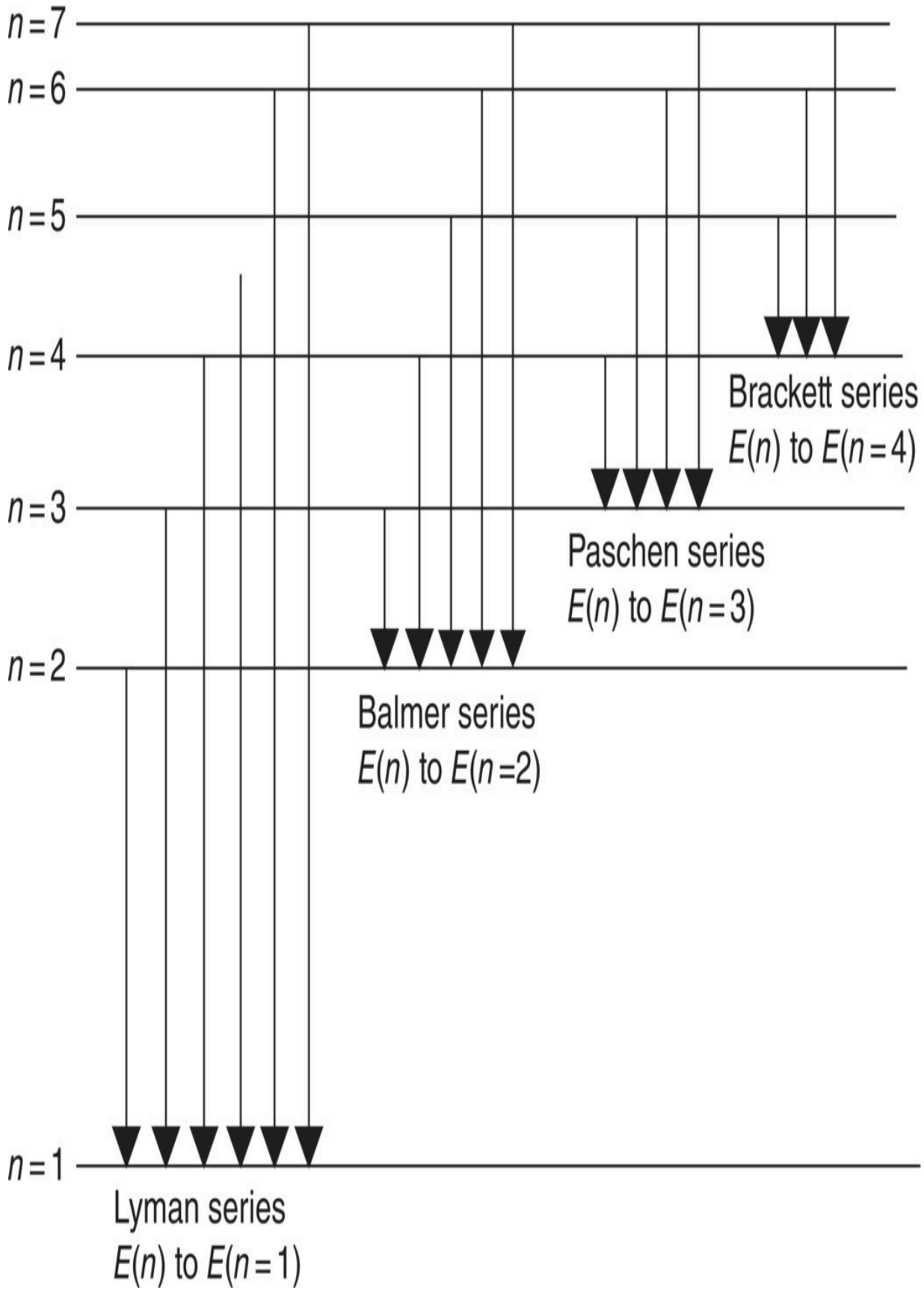

n=7
n=6
n=5
n=4
n=3
n=2
n=1
Brackett series
E(n) to E(n=4)
Paschen series
E(n) to E(n=3)
Balmer series
E(n) to E(n=2)
Lyman series
E(n) to E(n=1)

Figure 1.14 The observed energy lines of the hydrogen atom corresponding to all the transitions between different atomic levels.

In the next chapter, I discuss how these single unique energy levels that Bohr postulated explain the electric properties of different materials.

1.9 Summary and Conclusions

Perhaps the best way to summarize what we have covered in this chapter is to take a look at Figure 1.15. Observations of the sun's light spectrum and the spectra of different gases with their distinct lines resulted in heuristic relationships that relate the frequency of the missing lines to an expression consisting of just a constant and integer numbers. Einstein, working with the photoelectric effect, postulated the dual nature of light acting as both a wave and a particle, which we now call the photon.

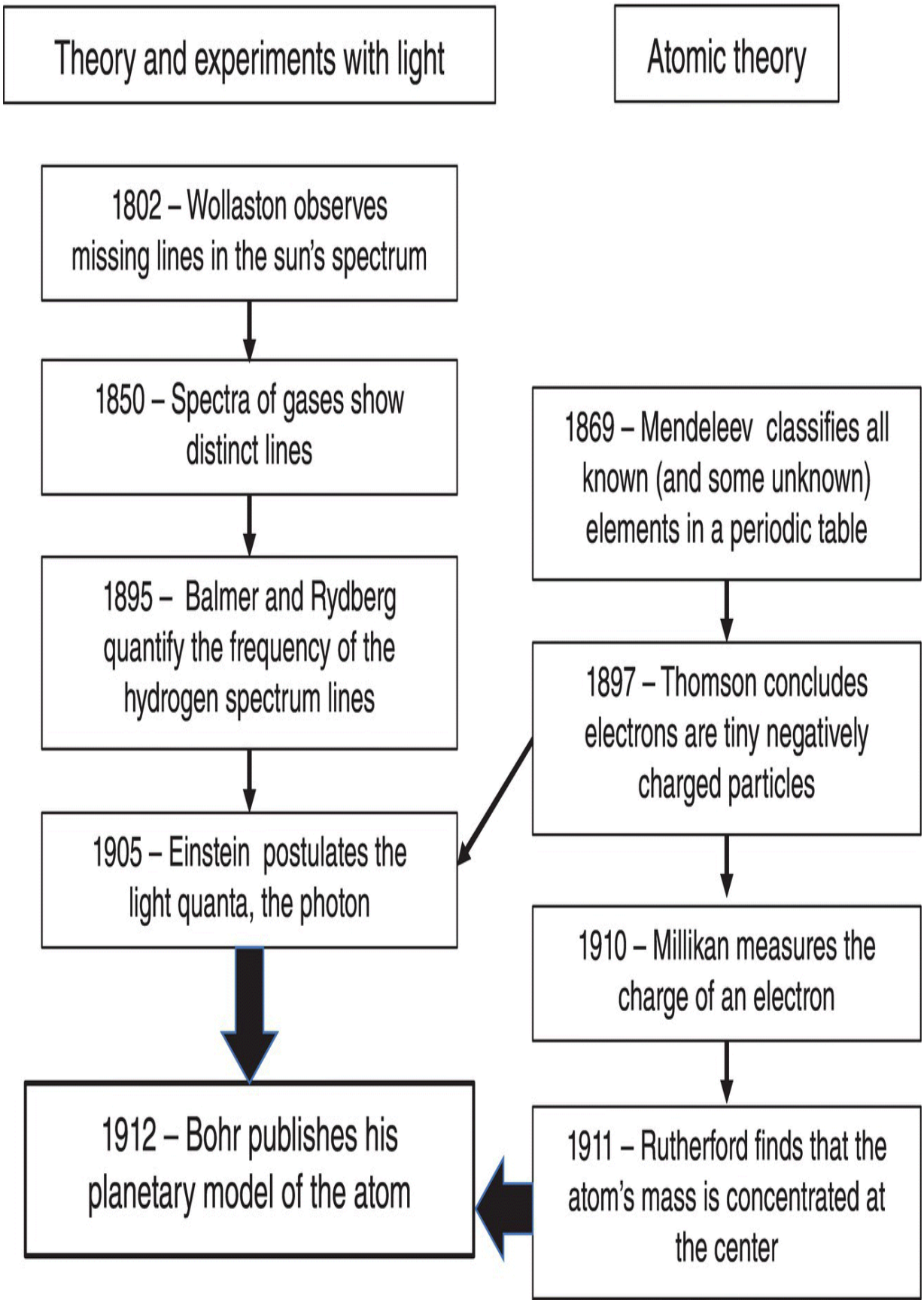

Theory and experiments with light
Atomic theory
1802 – Wollaston observes missing lines in the sun's spectrum
1850 – Spectra of gases show distinct lines
1895 – Balmer and Rydberg quantify the frequency of the hydrogen spectrum lines
1905 – Einstein postulates the light quanta, the photon
1869 – Mendeleev classifies all known (and some unknown) elements in a periodic table
1897 – Thomson concludes electrons are tiny negatively charged particles
1910 – Millikan measures the charge of an electron
1911 – Rutherford finds that the atom's mass is concentrated at the center
1912 – Bohr publishes his planetary model of the atom

Figure 1.15 The scientific and experimental work that led to the Bohr planetary model of the atom.

On the atomic side, Mendeleev classified the known elements by their weight, and, in the process, he left some empty spaces to add future elements. Thomson determined that electrons are tiny negatively charged particles and Rutherford, with his alpha ray measurements, concluded that the nucleus is concentrated in a very small region at the center of the atom. Millikan was able to measure the charge of electrons.

Based on previous theoretical and experimental work, Bohr proposed his planetary model of the atom with discrete and stable energy levels. His model included all the developments of atomic theory known to that date and explained the previous optical observations and measurements beautifully by considering how electrons move from one level to another by accepting or releasing packets of energy.

If you are comfortable with these conclusions, you are ready to go to the next chapter. You may peruse the three following appendices for a few more details.

Appendix 1.1 Some Details of the Bohr Model

Four quantum numbers uniquely define an electron:

The principal quantum number, *n*, defines the orbits and, therefore, the energy of the electrons. The energy released or absorbed as the electrons change orbits is determined exclusively by the value of *n*. The allowed values of *n* are any positive integers: 1, 2, 3, 4, etc.

Electrons have two spins: up and down.

Electrons also have angular momentum. The angular quantum number, ℓ, is associated with the shape of the orbits. The value of ℓ is also an integer number between 0 and *n*.

The magnetic quantum number, m_l, is associated with the orientation of the orbits. It is also an integer number between $-\ell$ and $+\ell$.

Electrons follow these rules:

The electrons in the first orbit, $n = 1$, can have two spins, but both ℓ and m_l are 0. Thus the first orbit can hold only two electrons.

The electrons in the second orbit, $n = 2$, can have two spins and two angular quantum numbers (ℓ equal to 0 or 1). Associated with $\ell = 0$, only one m_l value is possible, $m_l = 0$. But for $\ell = 1$, there are three possible values of m_l: -1, 0, and $+1$. So, the total number of electrons in the second orbit is eight: that is, four m_l times two spins each.

You can do the calculations and prove that the third orbit can hold up to 18 electrons, and so on.

Figure 1.16 shows one way of visualizing these levels, including the relative energy of the orbits of the Bohr atom and the order in which the orbits are filled. At 0 K absolute temperature, the electrons first fill up the 1s band (two electrons), the next two electrons reside in the 2s band, and the next six are in the 2p band, etc., climbing up the energy level stairs. Note that level 4s fills up before 3d. (By the way, the letters mean s for sharp, p for principle, d for diffuse, and f for fundamental.)

Another point you may wonder about is why we write the 3d level at higher energy than the 4s level. There is quite a debate about explaining this. It has to do with the attraction and repulsion of electrons and protons, and I will leave it at that.

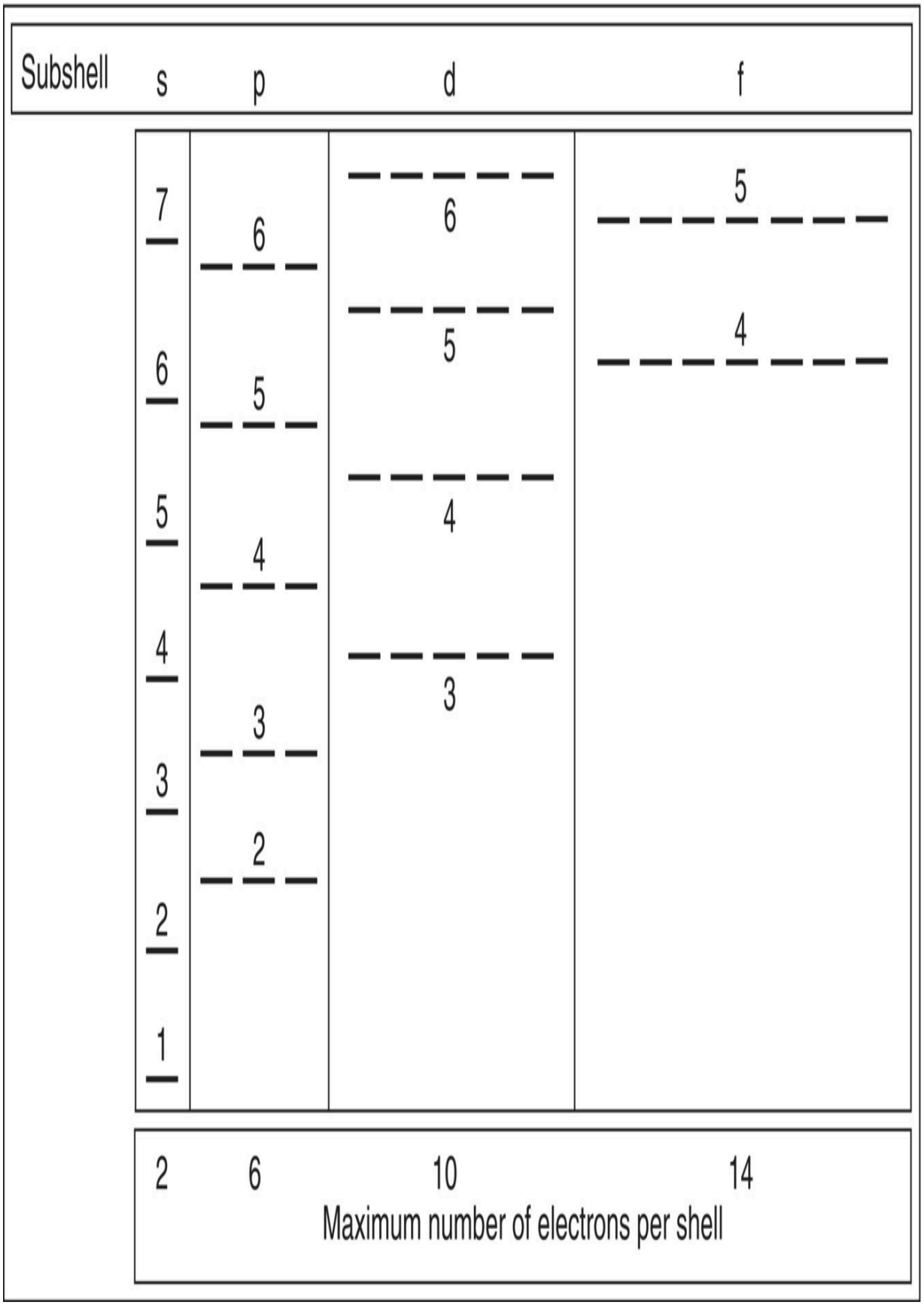

Subshell
s
p
d
f
7
6
5
4
3
2
1
6
5
4
3
2
6
5
4
3
5
4
2
6
10
14
Maximum number of electrons per shell

Figure 1.16 Subshell electron capacity. Notice that the number of sites in each level increases as the energy level, *n*, increases. Also notice that the 3d level has lower energy than the 4s level.

Appendix 1.2 Semiconductor Materials

Figure 1.17 shows the portion of the periodic table that contains the elements used in semiconductors, showing how many electrons are in the upper shells. All the lower shells are full. Silicon, for example, has 14 electrons, so the electrons fill up bands 1s (2), 2s (2), 2p (6), 3s (2), and 2p (the last 2). If you superimpose these numbers into Figure 1.16, you see that the two lowest energy levels are full, but energy level $n = 3$ (s + p) has four electrons, with the possibility of accepting another four (in the 2p level) to complete its orbit.

Consider another element that is used a great deal in semiconductors: antimony, Sb. It has 51 electrons. By looking again at Figure 1.16, we find the last occupied level is 5p, with three electrons. All the lower levels are full. Thus, the last occupied energy level, level 5, has five electrons – two in 5s and three in 5p levels – which gives it a chemical valence of 5. We will use these numbers in the next chapter to explain the difference between insulators, conductors, and semiconductors.

Group	II	III	IV	V	VI
2		5 BORON $2s^2 2p^1$	6 CARBON $2s^2 2p^2$	7 NITROGEN $2s^2 2p^3$	8 OXYGEN $2s^2 2p^4$
3		13 ALUMINUM $3s^2 3p^1$	14 SILICON $3s^2 3p^2$	15 PHOSPHORUS $3s^2 3p^3$	16 SULFUR $3s^2 3p^4$
4	30 ZINC $4s^2 3d^{10}$	31 GALIUM $4s^2 3d^{10} 4p^1$	32 GERMANIUM $4s^2 3d^{10} 4p^2$	33 ARSENIC $4s^2 3d^{10} 4p^3$	34 SELENIUM $4s^2 3d^{10} 4p^4$
5	48 CADMIUM $5s^2 4d^{10}$	49 INDIUM $5s^2 4d^{10} 5p^1$	50 TIN $5s^2 4d^{10} 5p^2$	51 ANTIMONY $5s^2 4d^{10} 5p^3$	52 TELURIUM $5s^2 4d^{10} 5p^4$
6	80 MERCURY $6s^2 4f^{14} 5d^{10}$	81 TITANIUM $6s^2 4f^{14} 5d^{10} 6p^1$	82 LEAD $6s^2 4f^{14} 5d^{10} 6p^2$	83 BISMUTH $6s^2 4f^{14} 5d^{10} 6p^3$	84 POLONIUM $6s^2 4f^{14} 5d^{10} 6p^4$

Figure 1.17 Portion of the periodic table emphasizing elements used in semiconductors.

Appendix 1.3 Calculating the Rydberg Constant

To complete some of the details of this chapter, the Rydberg constant can be calculated from more basic units. It is

$$R = \frac{m_e e^4}{8\varepsilon^2 h^3 c} = 1.1 \times 10^7 \,\mathrm{m}^{-1} \tag{1.6}$$

where m_e is the mass of the electron ($m_e = 9.1 \times 10^{-31}$ kg), e is the electronic charge ($e = 1.6 \times 10^{-19}$ C), ε is the permittivity of the material, that is the product of the relative permittivity of the material (e.g. for hydrogen $\varepsilon_r = 253.8$) times the permittivity of free space ($\varepsilon_0 = 8.85 \times 10^{-12}$ m^{-3} kg^{-1} s^2 C^2), h is Planck's constant ($h = 1.06 \times 10^{-34}$ J s), and c is the speed of light ($c = 3 \times 10^8$ m s^{-1}). All of these quantities are fundamental physical constants.

One thing that we often forget or do not check is the fact that any measurement is composed of a numerical value and a unit. In Eq. (1.6) I show several quantities, each of which has a number and unit(s). The result has to agree with the numerical calculation and the units. Very few people check the units when they perform an operation, which can result in mistakes (remember the fiasco when the Mars orbiter failed because of the confusion between metric and English units). So, let me do this with Eq. (1.6). First the units:

$$\text{units} = \frac{\text{kg} \times \text{Q}^4}{\left(\text{m}^{-3}\text{kg}^{-1}\text{s}^2\text{Q}^2\right)^2 \left(\text{kg} \times \text{m}^2\text{s}^{-1}\right)^3 \text{ms}^{-1}} = \frac{\text{kg} \times \text{Q}^4}{\text{m} \times \text{kg} \times \text{Q}^4} = \frac{1}{\text{m}} = \text{m}^{-1}$$

Look at the three terms in the denominator. There are meter (m) terms in the denominator with exponents −6 + 6 + 1 = 1 m, so only one meter unit remains in the denominator. Similarly, with the exponents of the kilograms, there are −2 + 3 = 1 in the denominator. There are four Qs in the denominator, and the seconds in the denominator cancel out (4 – 3 – 1 = 0). Now the kilograms and the coulombs in the numerator cancel those in the denominator, leaving only the reciprocal of a meter as the remaining unit, in agreement with Eq. (1.6).

Now let's do the numbers.

$$R = \frac{9.1 \times 10^{-31} \times \left(1.6 \times 10^{-19}\right)^4}{8 \times \left(8.85 \times 10^{-12}\right)^2 \times \left(6.63 \times 10^{-34}\right)^3 \times 3 \times 10^8}$$

$$= \frac{9.1 \times 10^{-31} \times 6.55 \times 10^{-76}}{8 \times 78.32 \times 10^{-24} \times 291.4 \times 10^{-102} \times 3 \times 10^8} = \frac{59.6 \times 10^{-107}}{547700 \times 10^{-118}}$$

$$= \frac{5.96 \times 10^{-106}}{5.48 \times 10^{-113}} = 1.1 \times 10^7$$

This agrees with the published result.

2
Energy Bands

OBJECTIVES OF THIS CHAPTER

We saw in the previous chapter that an atom's electrons have precise energy values (we represent them as orbits or levels). We also saw that electrons must have distinct quantum numbers (designations), which limits the number of electrons in an atom that can have a given energy. As atoms have more and more electrons, the electrons have to occupy higher and higher energy levels. An electron must absorb from somewhere the exact energy needed to jump from one level to a higher one. When it falls back to a lower level, it donates the same amount of energy. This is what happens in a gaseous state where the atoms are separated by large distances and do not interact with each other. This model beautifully explains the absorption and emission spectral lines of the elements, the sun, and the stars.

In this chapter, we are going to push the atoms closer and closer together until we form a solid. Now the atoms and electrons start interacting with each other and forming bonds, which is what keeps them together in a crystallographic structure. As we push them together, the energy levels have to separate because, in a system, according to Pauli's exclusion principle, no two electrons can have the same quantum number. The levels split into bands. Depending on how the bands spread, the material behaves like a conductor, an insulator, or a semiconductor.

Finally, we will analyze the specific case of semiconductors and how the electrons fit into the bands. We will also see how the lack of an electron, which we call a *hole*, is equivalent to a positive charge.

2.1 Bringing Atoms Together

From the previous chapter, we know that an atom has discrete energy levels, and electrons – unless disturbed by a packet of

energy – stay in a stable orbit. We show these energy levels as lines, as in Figure 2.1.

We also saw in the previous chapter that the Pauli exclusion principle states that no two electrons can share the same quantum numbers. If we push together two hydrogen atoms so that they interact and form a single system, the energy levels of the electrons in one atom must have slightly different values than the electrons in the other atom. I show this in Figure 2.2. What was a single level in hydrogen gas, with all the atoms separated by very large distances and acting as independent systems, now becomes two slightly different levels – yes, they are very, very close to each other, but they are still different. This ensures that all the electrons in the system, which is now composed of two atoms, have different quantum numbers.

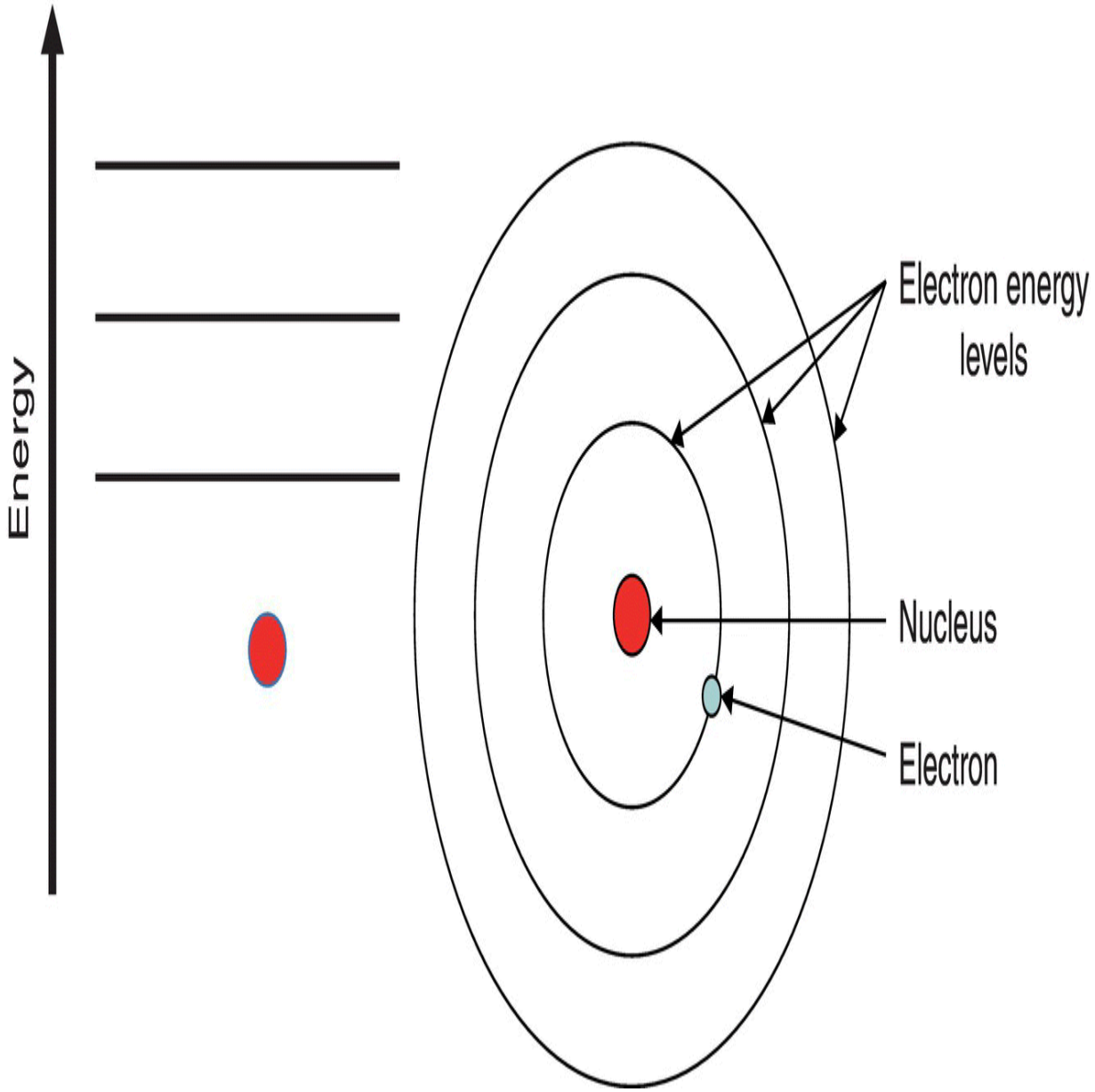


Figure 2.1 Energy levels in a Bohr atom (left) corresponding to the Bohr energy orbits (right).

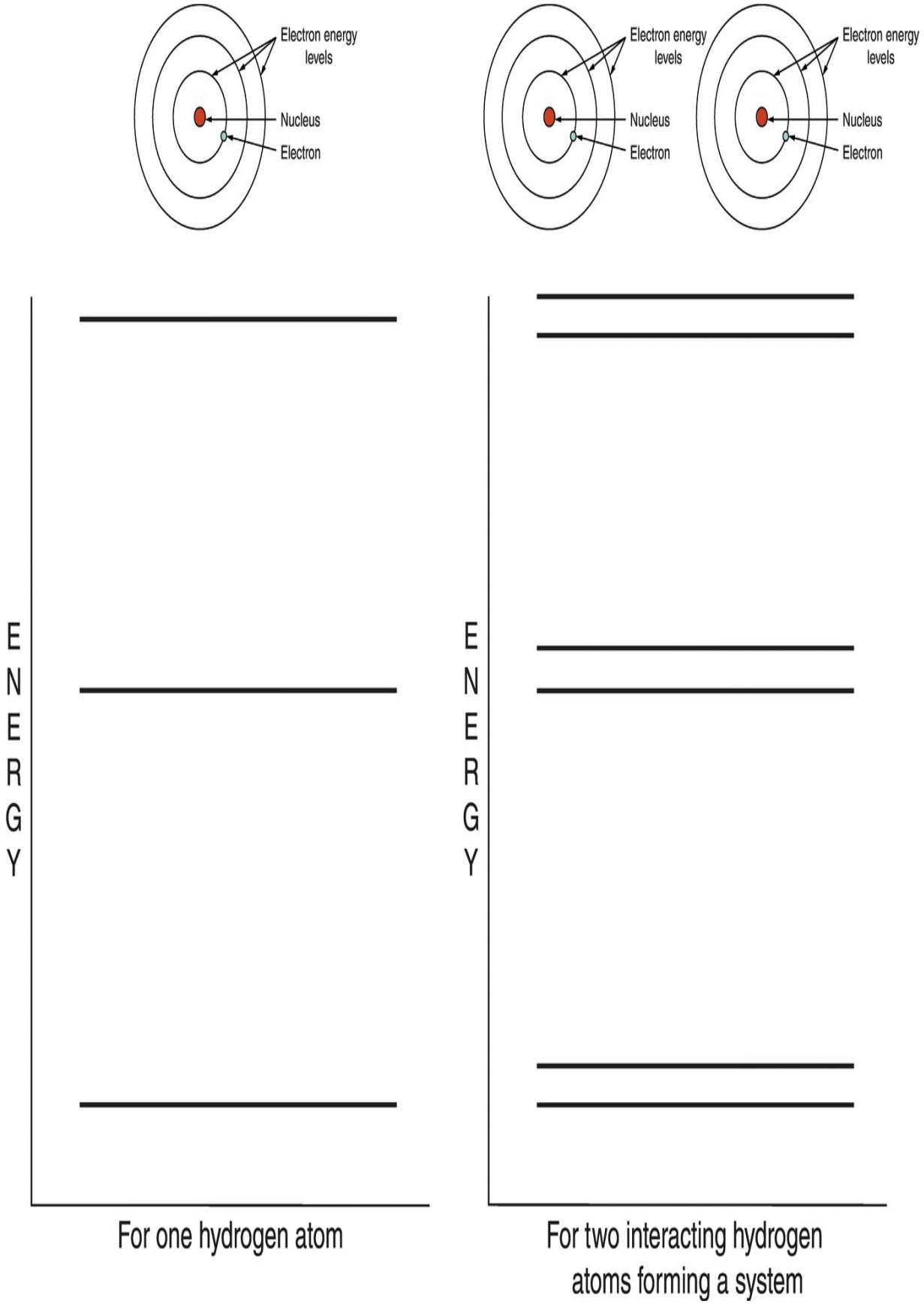

Electron energy levels
Nucleus
Electron
Electron energy levels
Nucleus
Electron
Electron energy levels
Nucleus
Electron
E N E R G Y
E N E R G Y
For one hydrogen atom
For two interacting hydrogen atoms forming a system

Figure 2.2 When two hydrogen atoms are so close that they form a single system (right), the Pauli exclusion principle requires that their energy levels be different.

Now consider the case of a solid. There are approximately 5×10^{22} silicon atoms per cm^3, that is, a 5 followed by 22 zeros in a space a little smaller than a sugar cube. We can no longer talk about discrete energy levels, as we did with a single hydrogen atom or hydrogen gas. Now we need to talk about energy bands (Figure 2.3). (Try to draw 10^{22} lines on a piece of paper!)

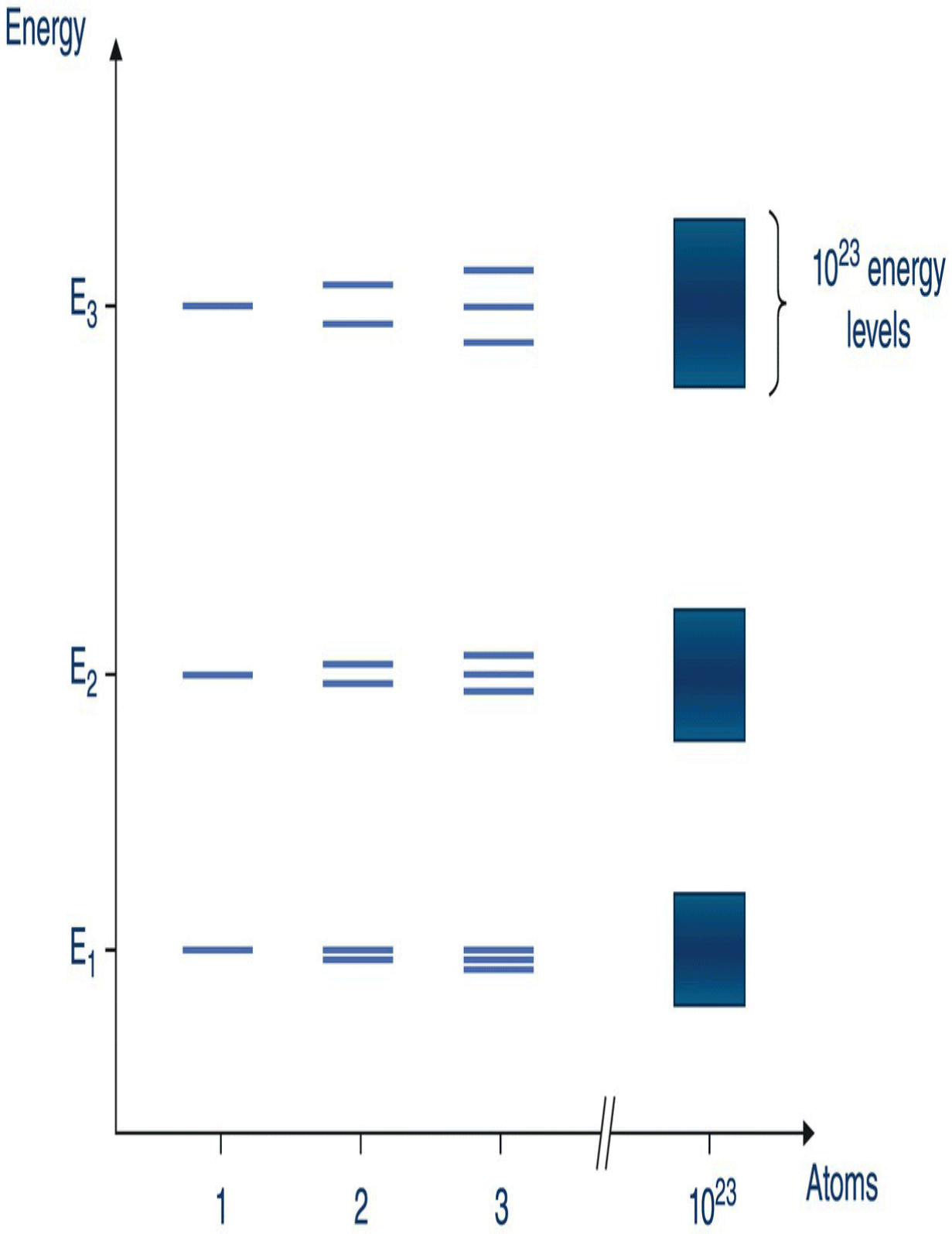

Energy
E_3
E_2
E_1
10^{23} energy levels
1
2
3
10^{23}
Atoms

Figure 2.3 From energy levels in a gas where the electrons in the atoms are separated (left) to energy bands in a solid where the atoms are in such proximity that they interact with each other (right).

Energy bands explain the difference between an insulator, a metal, and a semiconductor. Everything depends on how many electrons are in the atom's outer energy level, that is, the valence band level, how the atoms share electrons when they form a solid, and what the separation is between the bands.

Imagine a chamber filled with a gas. The atoms are bouncing around without interfering with each other. Now we start shrinking the chamber, making the distance between atoms – the interatomic distance, *a* – smaller and smaller (Figure 2.4).

When *a*, the interatomic distance, is very large (infinity) at the right of Figure 2.4, the atoms are so far separated that the electrons do not interact with each other. When they are that separated, all the individual atoms have the same energy levels. As the distance between atoms gets smaller, at some point the atoms (beginning with those in the outer orbits) start interacting, and the energy levels begin broadening to satisfy Pauli's exclusion principle. Nature determines the actual interatomic distance, *a*, in a solid. In silicon, for example, the interatomic distance (called the *lattice constant* in crystals) is 5.431 angstroms (5.431 Å: an A with a tiny o on top). An angstrom is equal to 1×10^{-10} m.

Note in Figure 2.4 that as the interatomic distance gets smaller, the levels start separating in different ways. At the solid's equilibrium spacing, the separation between the first and second bands, A and B, is very large; between the second and third levels, B and C, there is a separation, but it is much smaller. Finally, there is no separation between the third, fourth, and fifth bands (C, D, and E). They encroach on each other. This is expected, because the electrons in the outer bands start to interact before the inner ones do.

At absolute zero (0 K, –273°C, –460°F), there is no energy whatsoever in the system, and all electrons are at the lowest allowed orbit; in a solid, they are all in the lowest possible energy band. We call the highest energy band occupied with some electrons at 0 K the *valence* band, and we call the next energy band above it – which is *totally* empty of electrons – the *conduction* band. You'll see why in a minute. We call the energy separation between the two bands the *energy gap* (*Eg*) because no electrons are allowed to have these in-between energies. Here are all the possible cases at absolute zero:

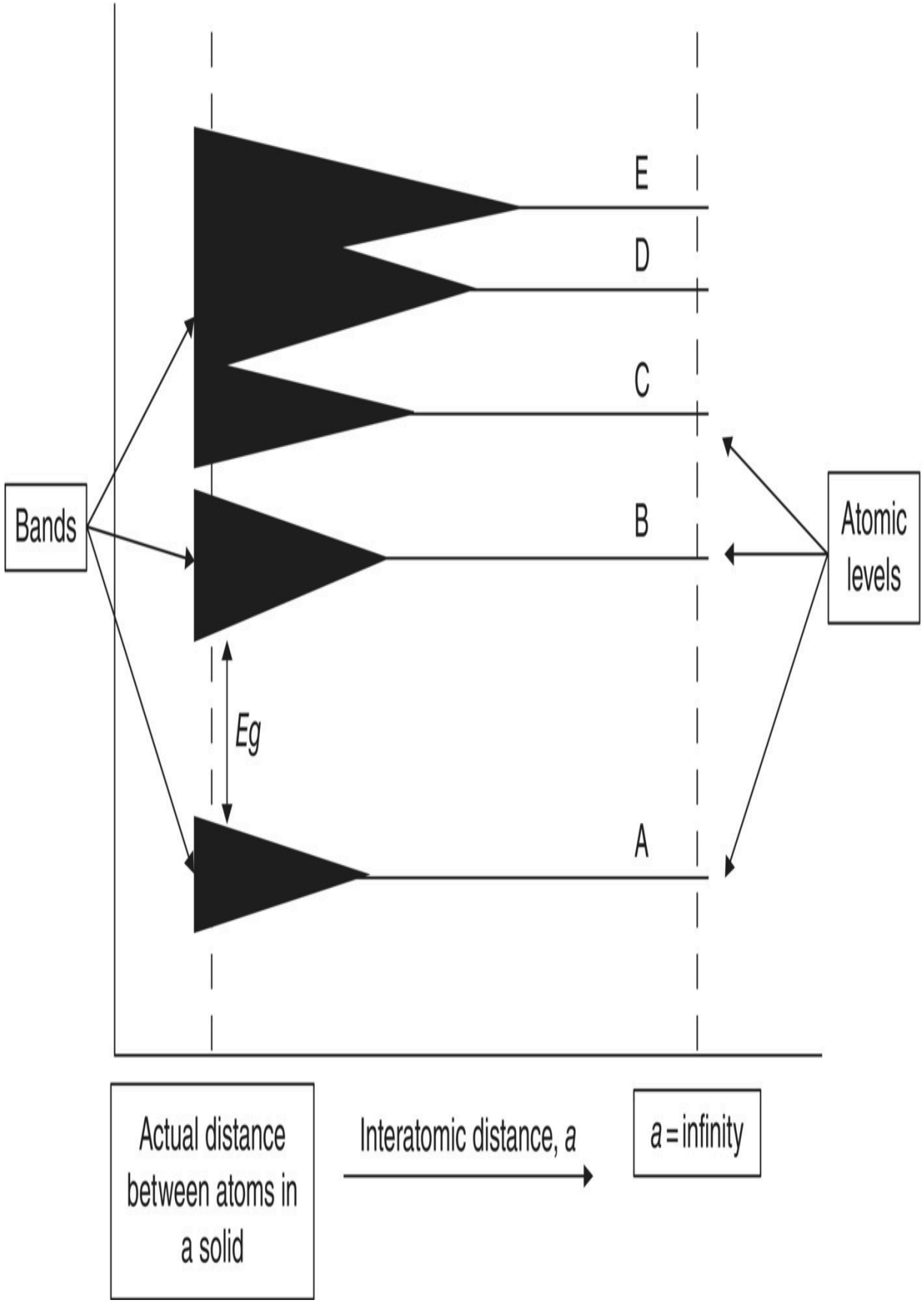

E
D
C
B
A
Bands
Atomic levels
Eg
Actual distance between atoms in a solid
Interatomic distance, a
a = infinity

Figure 2.4 Atomic levels split into bands as the interatomic distance between atoms becomes smaller, forming energy bands and energy gaps as soon as they are close enough to interact with each other.

The valence band is full of electrons; there is no space for any more.

The valence band is not full of electrons. There are empty spaces that can accept additional electrons.

The conduction band is separated from the valence band with a small energy gap, such as the separation of levels B and C.

The conduction band is separated from the valence band with a large energy gap, such as that between A and B.

There is no separation at all between two bands such as C and D or D and E. The conduction band touches or falls inside the valence band.

Let's take a look at each of these cases.

2.2 The Insulator

Let's consider first the situation in which the valence band is full of electrons (case 1) and there is a large gap separating the valence and conduction bands (case 4). I illustrate this in Figure 2.5. Imagine that the valence band is a parking lot full of cars (electrons), and above it is the conduction band, equivalent to an empty freeway. The cars in the parking lot cannot move because there is no space for them to go to. No cars are on the freeway, so there is no movement up there, either. If I apply a voltage to this material (right, in Figure 2.5), nothing moves. The gap between the parking lot and the freeway is too large for cars to jump, and thus no matter what voltage I apply, the electrons cannot move; there is no current flowing in this material at all. This is the case for *insulators*. No electrons can move under an applied voltage.

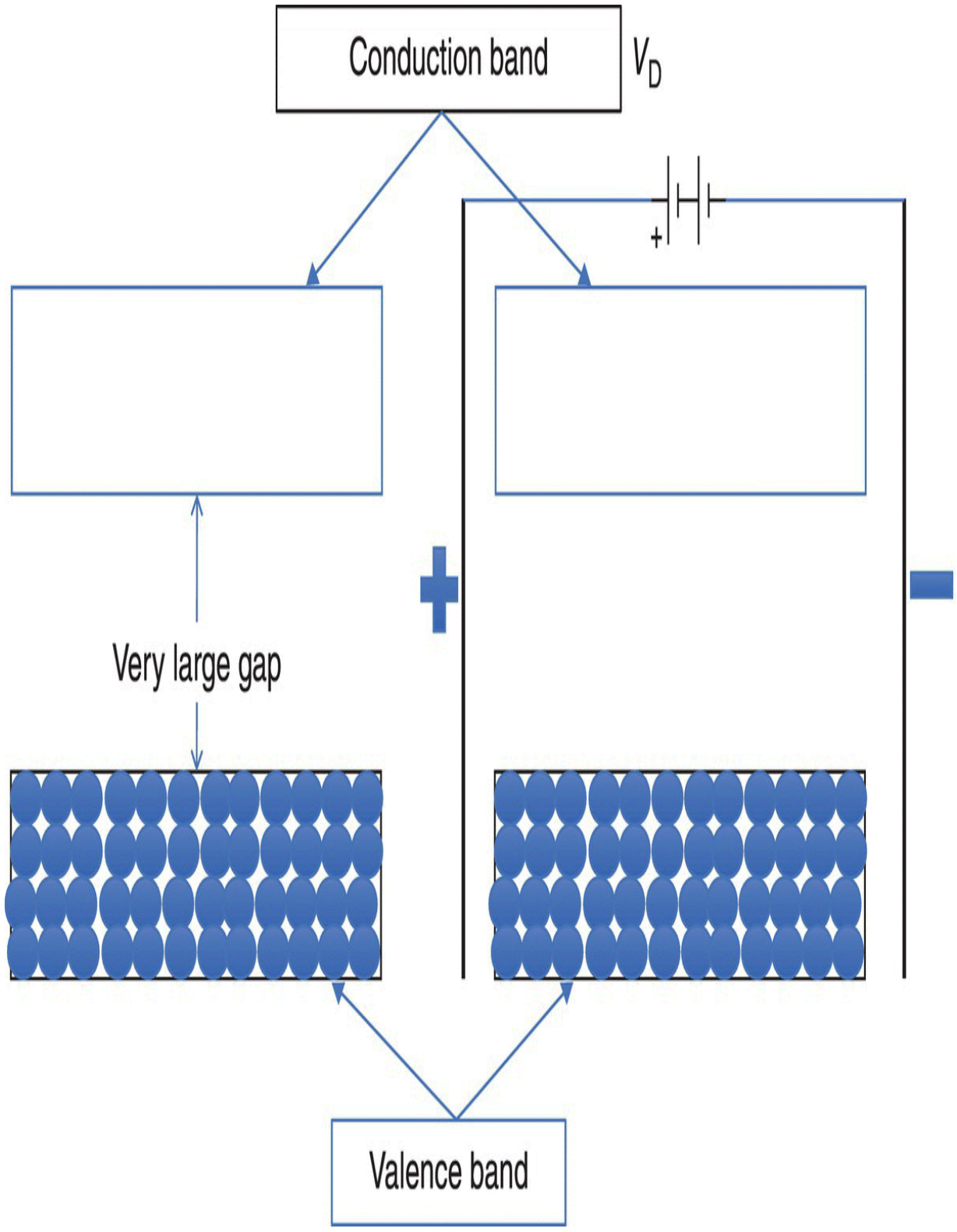

Conduction band
V_D
+
Very large gap
Valence band

Figure 2.5 In an insulator, the valence band is full of electrons, the conduction band is empty, and the separation between the two bands is very large.

At this point, I would like to clarify the concept of bands, which can sometimes be confusing. The bands are not a physical location where electrons reside, just as earth's orbit is not a railroad track or a circular road on top of which the earth travels. A cannonball follows a parabolic path even though there is no "path," pipe, road, or track that the ball rolls over. The concept of energy bands is similar. The electrons are anywhere in the material, but they have energies with values restricted to certain allowed ranges – energy ranges that we call *bands*. The electrons are not allowed, under any circumstances, to have energy between the highest energy of the valence band and the lowest energy of the conduction band.

2.3 The Conductor

Now consider the situation where the valence band is not full of electrons (case 2), as I show in Figure 2.6, or where the bands expand so much that the conduction band encroaches into the valence band (case 5), as I show in Figure 2.7. These two cases are very similar.

Even at absolute zero (as I show on the left), there is lots of space for the electrons to move. At room temperature, which is quite an increase in energy from absolute zero, the electrons have more than enough energy to move all over the place. It is like having a half-full garage with two doors at each end connected directly to a freeway, and it is warm enough that people want to get moving. If we apply a voltage (the motivating force) across this material, the electrons have no problem going where they want to go, i.e. to the positive terminal. Also note that the cars that moved up to the freeway left empty spaces in the garage, so some cars in the garage can also move from one location to another inside the parking structure. As you may expect, this is a *conductor*. Tons of electrons are free to move as soon as a voltage is applied.

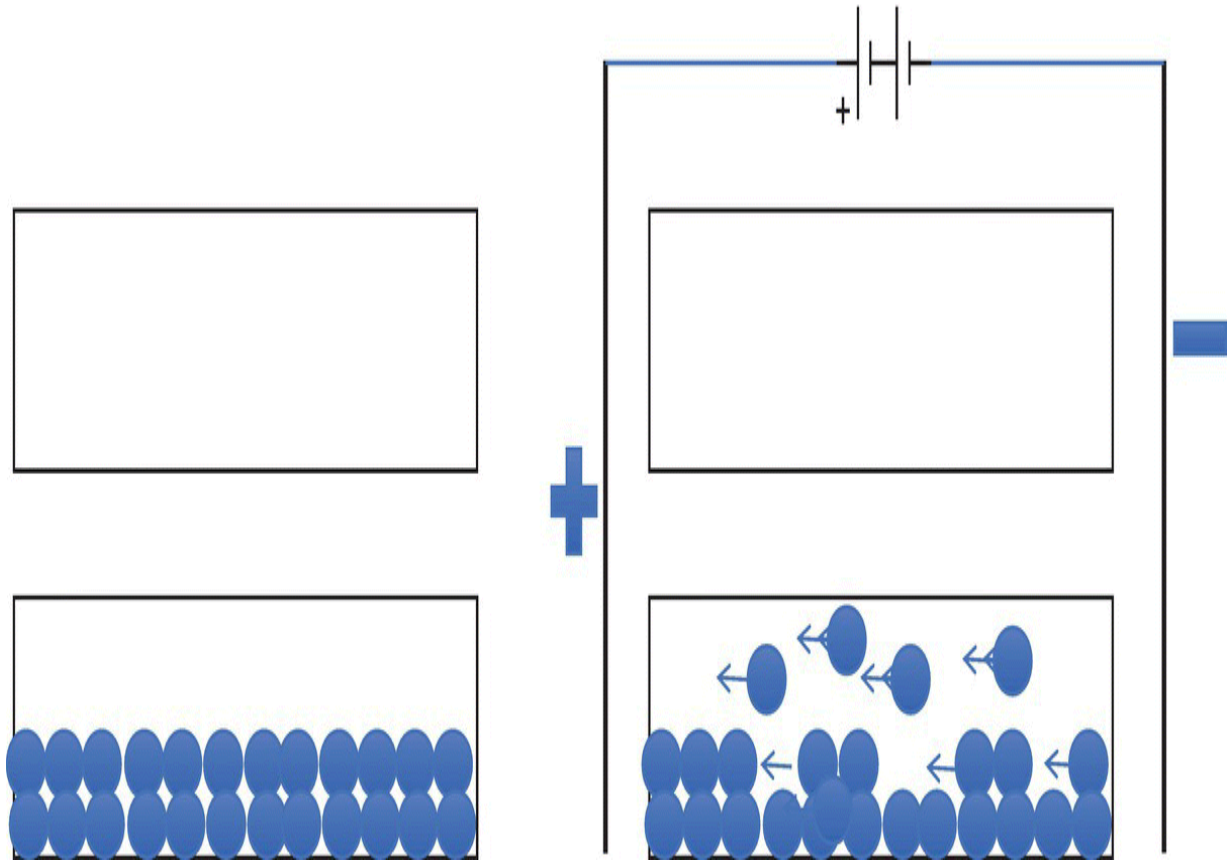


Figure 2.6 If the valence band is not full of electrons, there is a lot of space even at the lowest temperatures for electrons to move easily in the valence band.

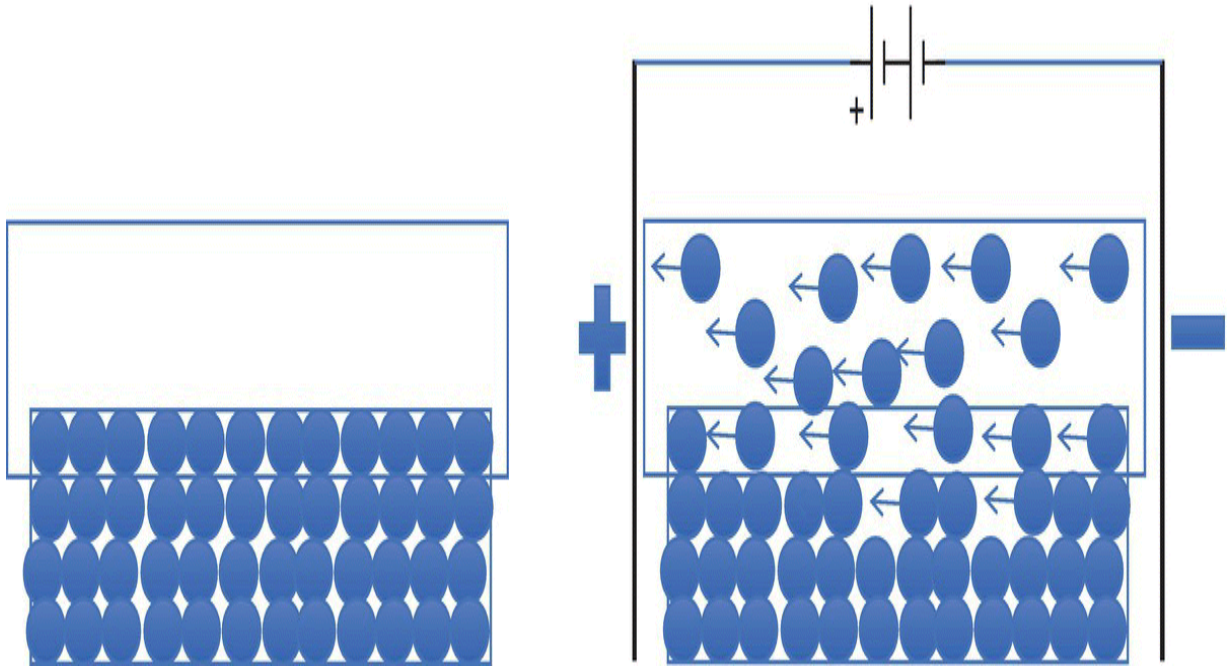


Figure 2.7 Even if the valence band is full, if the conduction band encroaches on the valence band, there is plenty of space for the electrons to move freely when a force (a voltage) is applied.

2.4 The Semiconductor

Now consider the situation where the valence band is full (case 1), but there is a small gap between the conduction and valence bands (case 3). At absolute zero, all the electrons are in the valence band, the lowest allowed energy, and thus the valence band is completely full. But at room temperature, there is enough thermal energy in the system that, statistically, a few electrons can make the jump from the valence band to the conduction band. I show this in Figure 2.8.

In pure silicon, as I said earlier, there are 5×10^{22} silicon atoms per cm^3; but only a tiny number (1.45×10^{10} electrons per cm^3) have, statistically, sufficient energy at room temperature to jump from the valence band to the conduction band. Although 10^{10} electrons per cm^3 looks very large, this is only a single solitary electron out of 3.4×10^{12} silicon atoms (1 out of 3 400 000 000 000) that gain the energy needed to jump to the conduction band. This situation is exemplified in Figure 2.8. Note that as the very few electrons jump to the conduction band, they leave behind an empty spot, which we call a *hole*. The holes are called *p-particles* (*p* for *positive*) because they appear to have a positive charge. (Remember, each silicon atom is neutral, that is, it has as many protons – positive charges – in the nucleus as it has electrons in the orbits. If you remove one electron, the atom becomes positively charged.) Notice also that in an intrinsic (i.e. perfectly pure) semiconductor, the number of electrons, n_i, and the number of holes, p_i, have to be exactly the same, since every electron that jumps to the conduction band leaves behind an empty space (a hole) in the valence band. We refer to these intrinsic electrons and holes using the subscript i.

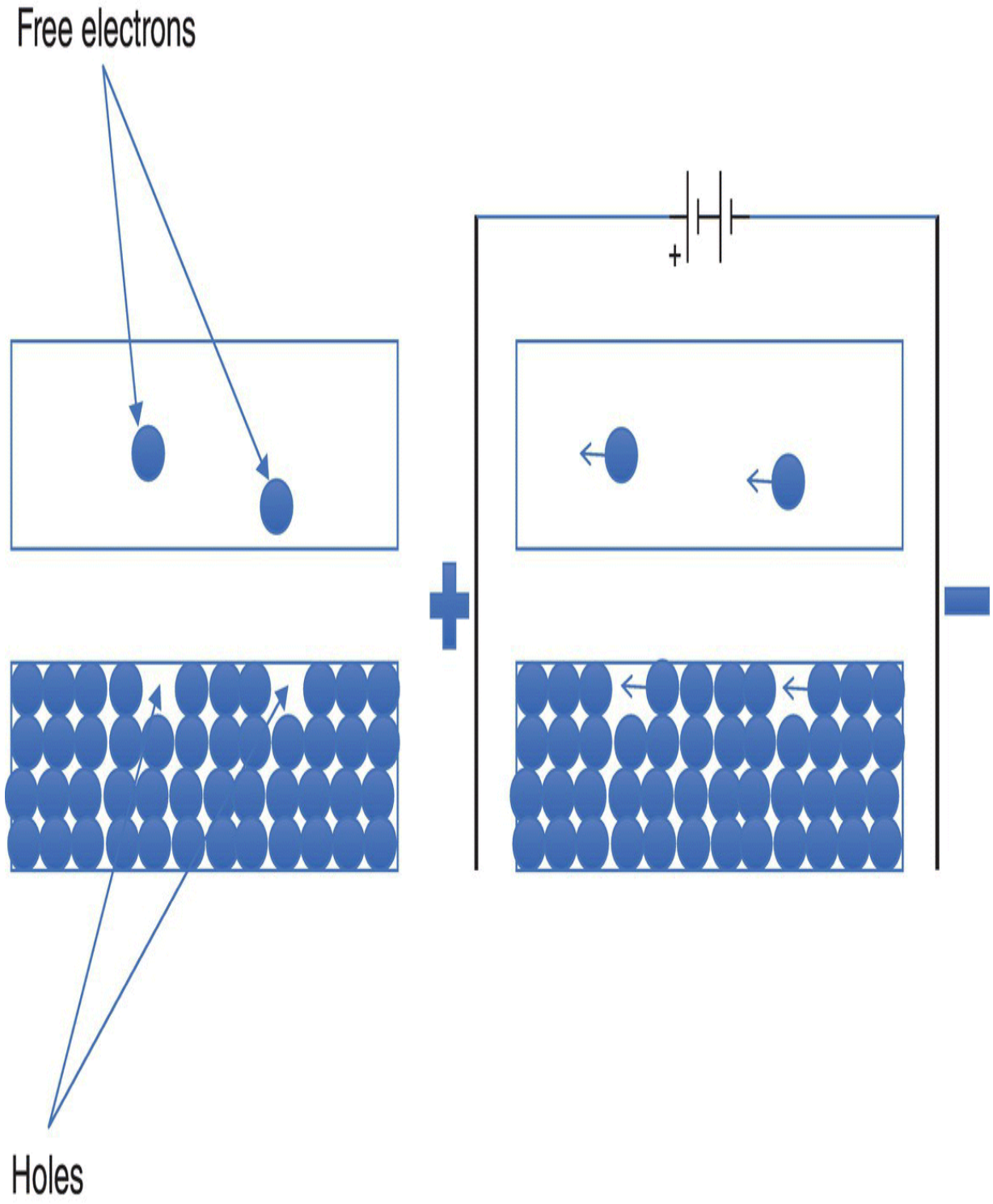

Free electrons
+
Holes

Figure 2.8 The valence band in a semiconductor is completely full, the conduction band is empty, and the separation between the two bands is small, but at room temperature, which is what I show here, there is sufficient energy to kick a few electrons from the valence band to the conduction band, thus generating a small current when we apply a voltage.

Consider what happens when we apply a voltage, as I show on the right in Figure 2.8. Electrons move left toward the positive side: both the electrons in the conduction band and those electrons in the valence band that have an empty space to their left. We like to say, and you will see why later, that the hole (the empty space) is moving in the opposite direction – to the right, toward the negative side – so the hole acts like a positive particle moving to the negative terminal.

Since it is the thermal energy that determines how many electrons can jump to the conduction band and are free to move, and how many holes are left behind, you can expect that the current in a semiconductor is highly dependent on the temperature, that is, the thermal energy. Figure 2.9 shows the number of free electrons in silicon and gallium-arsenide (GaAs) in the conduction band as a function of temperature.

Note that the number of intrinsic electrons and holes in the plot (the vertical *y* axis) is logarithmic, and the temperature (*T*, on the horizontal *x*-axis) is not the temperature but 1000/temperature in absolute units (Kelvin). The actual temperature in degrees Celsius is at the *right side* of the plot. The advantage of plotting it this way is that the change in the number of free electrons versus $1/T$ is very linear. For convenience, I mark the values at room temperature (27 °C), at freezing (0 °C) and at the boiling point of water (100 °C).

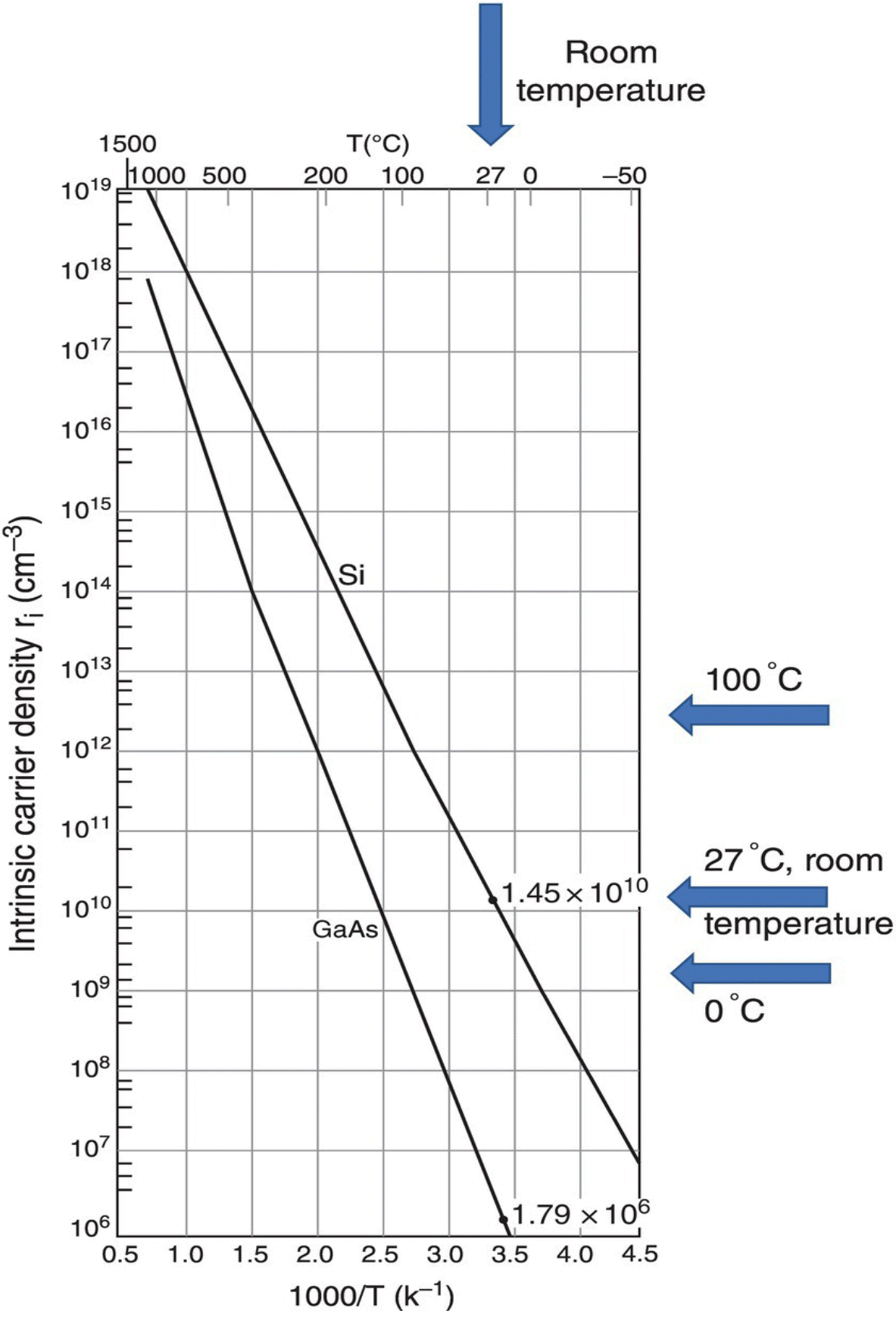

Room temperature
1500
T(°C)
1000 500 200 100 27 0 –50
10^{19}
10^{18}
10^{17}
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
Intrinsic carrier density r_i (cm^{-3})
Si
GaAs
1.45×10^{10}
1.79×10^{6}
100 °C
27 °C, room temperature
0 °C
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5
1000/T (k^{-1})

Figure 2.9 Electron and hole concentrations in Si and GaAs change drastically as a function of temperature. Temperature is an indication of the energy of the system, and the more energy in the system, the more electrons can move from the valence band to the conduction band, leaving behind the same number of holes.

The number of free charges is also dependent on the light. Photons can hit the silicon crystal and give their energy to an electron that can use this energy to jump to the conduction band.

Looking at Figure 2.9, at room temperature (27 °C, 300 K), the number of electrons and holes in silicon is $1.45 \times 10^{10}\,\text{cm}^{-3}$, as I said earlier; but if we cool it down to 0 °C (the temperature of ice water), the number decreases rapidly to $2 \times 10^{9}\,\text{cm}^{-3}$, a factor of 10 lower. If we do the opposite and immerse a semiconductor in boiling water (100 °C), the number of free charges increases to $3 \times 10^{12}\,\text{cm}^{-3}$, 300 times larger than at room temperature. As a rule of thumb, the number of free charges in silicon doubles every 7°C.

In the same graph, I have the number of free charges in GaAs, the next-most-used semiconductor. Because the separation between the valence and conduction bands (*Eg*) in GaAs is larger than that of Si (1.43 eV vs. 1.12 eV), the number of free carriers in pure GaAs material is much lower than that of silicon, $1.79 \times 10^{6}\ \text{cm}^{-3}$.

Remember this temperature dependence when we talk about the operation and use of semiconductor devices. When we turn on an electronic device, the temperature increases, and its electronic properties change. In Chapter 9, I discuss some of the tricks that designers use to stabilize circuits.

2.5 Digression: Water Analogy

Suppose you have a large pot of water sitting on the counter at room temperature (25 °C). The water is not moving. Now you place the pot on the stove and increase the temperature to 100 °C, the temperature of boiling water. There are three possibilities:

The rim of the pot is much higher than the level of the water. The boiling water remains in the pot. The rim is too high for the water to boil over.

The water fills the pot all the way to the top. As soon as the water starts boiling, water spill all over the stove.

Now consider an intermediate case. The water level is high, but it does not reach the rim. How much water spills over depends very much on the distance between the original water level and the rim of the pot. The larger the distance between the water level and the rim of the pot, the more water will spill from the pot.

You can immediately recognize the analogy. First, the separation of the water level and the rim of the pot is analogous to the energy gap. Case 1 is the insulator, where there is a large distance between the electrons and the empty conduction band. Case 2 is the conductor, where the electrons start spilling into the conduction band as soon as there is any energy at all. Finally, case 3 is the semiconductor, where the number of electrons in the conduction band depends on the separation between bands. There are more free electrons in silicon at room temperature than in GaAs because the separation of bands is much larger in GaAs.

I hope the analogy helps you understand the effect of the bands in defining the conductive processes of different materials. If you had difficulty in the previous sections, it might be worthwhile to go back and read them again.

2.6 The Mobility of Charges

Without going into detail, I would like to introduce the concept of mobility. *Mobility*, with the symbol μ, is basically a measure of how easy it is for electrons and holes to move: the opposite of resistance. Look at Figure 2.10. An electron in the conduction band without any impediments, like a car on the freeway, can move quite easily. The hole has more problems moving. Electrons in the valence band have to hop from one atom to an adjacent one that has empty space at

the left. The next electron in line can jump to the empty space that the previous electron left behind, and so on. To an external viewer, a positive charge is moving to the right. You can intuitively see that electron motion is easier than hole motion.

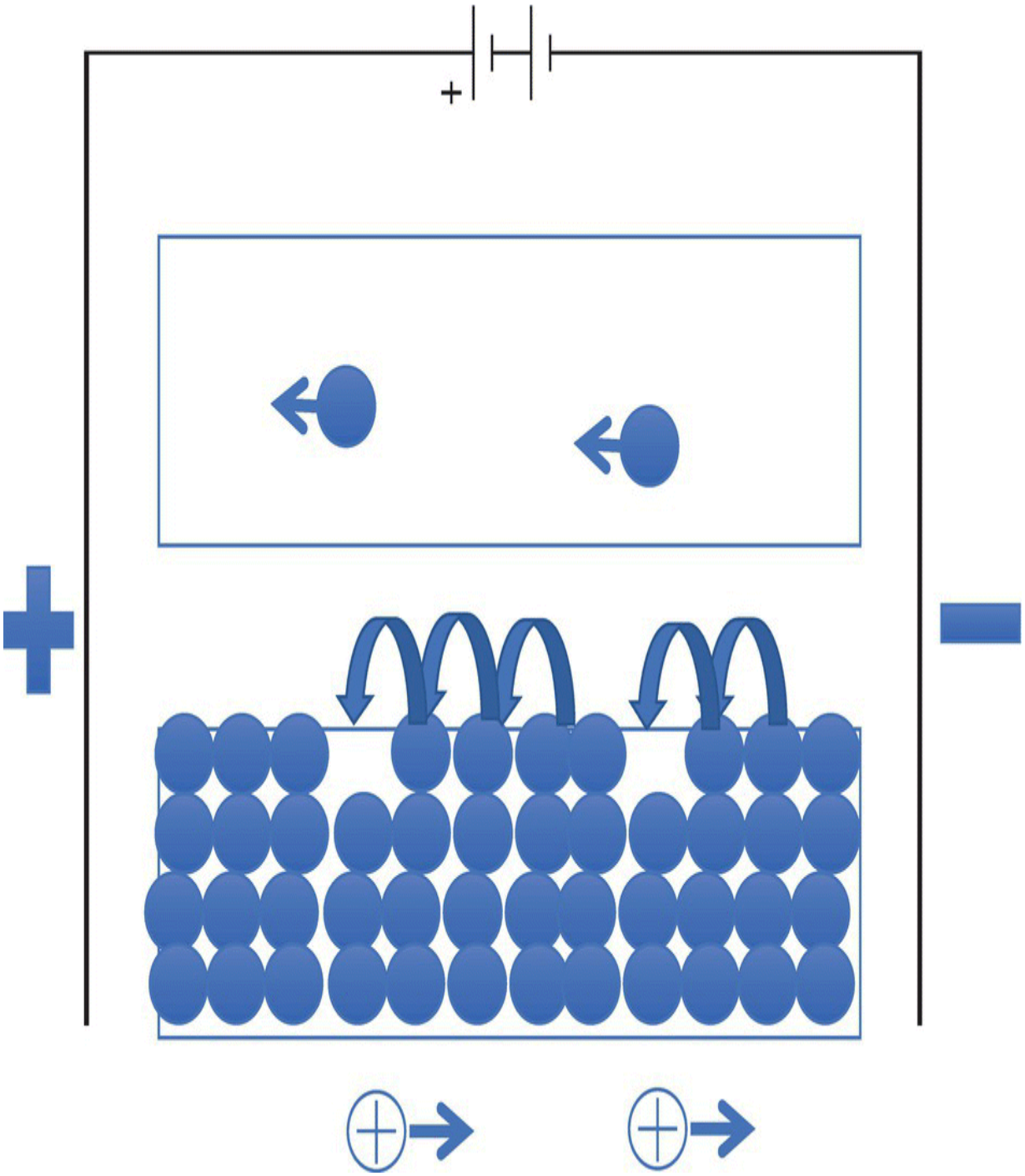


Figure 2.10 Electrons in the conduction band are free to move, while those in the valance band have to hop to the closest atom that is missing an electron, which means the mobility of electrons in the conduction band is higher than that of holes (electrons in the valence band).

The mobility of electrons in silicon is $\mu_n = 1400\,cm^2/(V \times s)$, while the mobility of holes is only $\mu_p = 450\,cm^2/(V \times s)$. For the same applied voltage, the electrons in silicon are more than three times faster than the holes. This makes a lot of sense, and it is true for all semiconductors.

2.7 Summary and Conclusions

In this chapter, we have seen how as we shrink the interatomic distance between atoms, due to the Pauli exclusion principle the gaseous energy levels spread and become bands. We call the lowest energy band that has electrons the *valance band* and the one above it, empty of electrons, the *conduction band*. Depending on how many electrons are in the valence band and how far the valence and conduction bands are separated, we have conductors, insulators, or semiconductors.

At absolute zero, the valence band of a semiconductor is completely full, and the conductions band is empty. The separation between the bands is small. At room temperature, there is quite a bit of thermal energy, and a very small number of electrons have sufficient energy to move to the conduction band – and they are free to move. In the valence band, the missing electrons leave a hole that acts as if it were a positive charge.

The key concepts in this chapter that you need to understand to proceed to the next chapter are as follows:

The concepts of a valence band and a conduction band.

The concept that a small energy gap is required for an electron to move to the conduction band.

The concept of the *hole* – an atom missing an electron, which, under an applied voltage, accepts an electron from a neighboring atom, thus moving the hole toward the negative terminal as if it were a positive charge.

If you are comfortable with these concepts, you are ready to go to the next chapter.

Appendix 2.1 Energy Gap in Semiconductors

Semiconductors have different energy gaps measured in electron-volts (eV), the energy of an electron under an applied voltage of 1 V (Figure 2.11). As you may expect, an electron-volt is a very small amount of energy ($e = 1.602 \times 10^{-19}$ coulombs). The energy gaps of the most commonly used semiconductors – germanium, silicon, and gallium arsenide – are 0.67, 1.12, and 1.43 eV, respectively. Silicon is the most commonly used because it is easier to grow as pure crystals with no (or an insignificant number of) impurities and imperfections, and it is also easier to inject a controlled number of impurities. In the next chapter, I discuss why purity and the ability to add a controlled number of impurities are so important.

Main semiconductor band gaps

Material	0 K	300 K
Si	1.17	1.11
Ge	0.74	0.66
GaAs	1.52	1.43
InSb	0.23	0.17
CdTe	1.61	1.44
InP	1.43	1.27

Figure 2.11 There is a large difference in energy gaps in semiconductors, from a very low value for InSb (0.17 eV) to GaAs (1.43 eV). Notice that the energy gap changes slightly as the temperature changes: about 5% for silicon between absolute zero and room temperature.

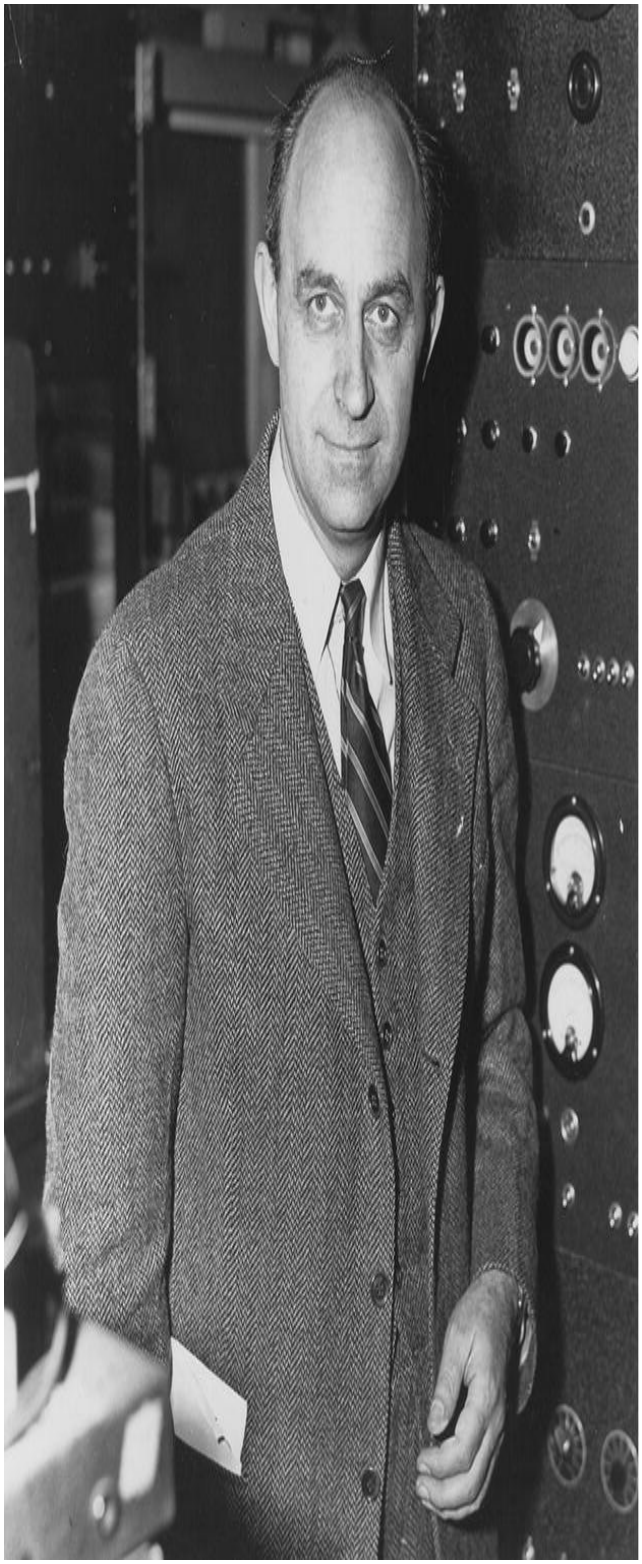
There is an interesting website with a huge amount of information on the elements: http://periodictable.com/Elements/050/data.html. Clicking any element in the periodic table will tell you its properties. Furthermore, if you click any property, it will tell you the value of that property for all the elements in the periodic table. It is interesting to see, using this website, that the great majority of elements are conductors. According to this site, of all the elements

in the periodic table, 76 are conductors, 5 are insulators, and only 3 are semiconductors: Si, Ge, and Te. Te has an energy gap of 0.45 eV. The rest of the elements are either gases or have unknown values.

Appendix 2.2 Number of Electrons and the Fermi Function

To help you better understand how many electrons and holes are free in a material, let me state, without a proof, the Fermi–Dirac function. Enrico Fermi (1901–1954) was an Italian physicist who emigrated to the United States in 1938 to escape the persecution of Jews (his wife was Jewish). Paul Dirac (1902–1984) was a British scientist. These two physicists did a huge amount of work on quantum theory (Figure 2.12).

For our purposes, we will look at the results of their quantum statistics: the Fermi–Dirac function, or F-D function for short, which both scientists developed independently in 1926.



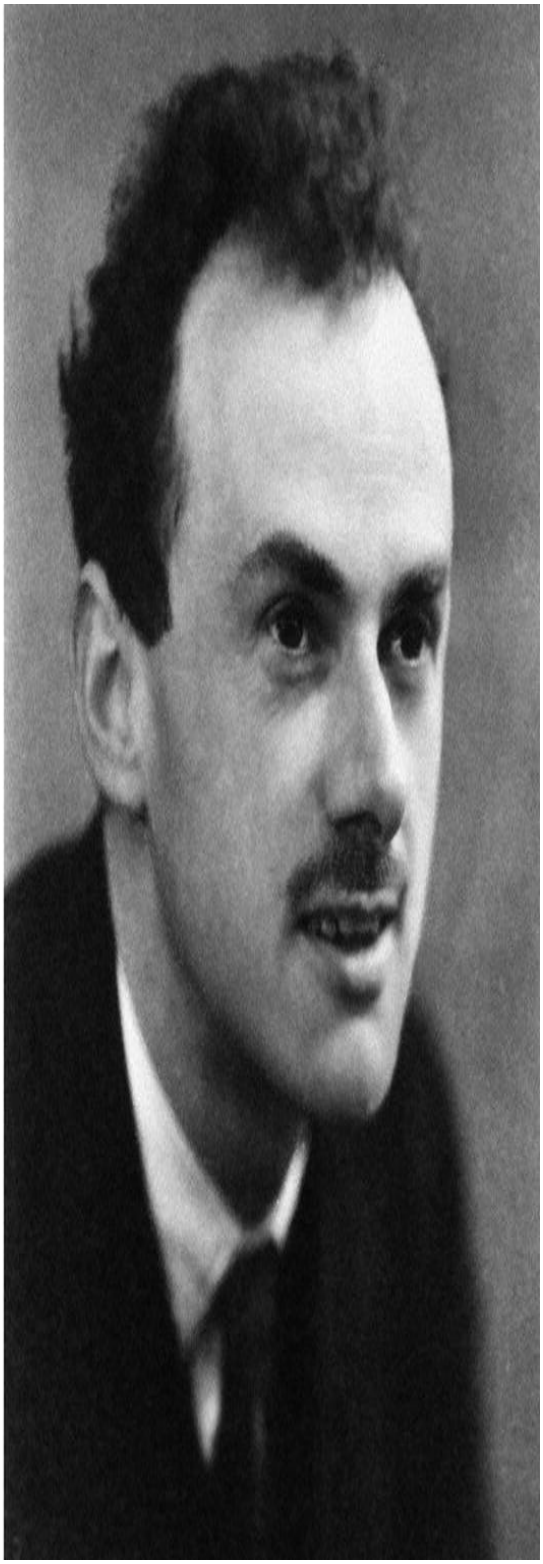


Figure 2.12 Enrico Fermi (left) and Paul Dirac (right), who developed the statistics for particles that obey the quantum physics theory.

Source: https://en.wikipedia.org/wiki/Enrico_Fermi#/media/File:Enrico_Fermi_1943-49.jpg (left); https://en.wikipedia.org/wiki/Paul_Dirac#/media/File:Dirac_4.jpg (right).

Ludwig Boltzmann (1844–1906), an Austrian physicist, developed his Boltzmann distribution function, which related the behavior of gases and their pressure to their temperature. This applied only to classical particles. The F-D equation applies the quantum mechanical theory to particles that behave under the Pauli exclusion principle. By the way, these statistics apply to all particles that obey the Pauli exclusion postulate, not just electrons, and these particles are called *fermions*.

Anyway, the F-D formula is

$$F(\mathrm{E}) = \frac{1}{e^{(E-E_\mathrm{f})/kT} + 1} \tag{2.1}$$

where F(E) is the probability that an energy level E is occupied by an electron, E is the energy of that specific level, E_f is the Fermi level, k is the Boltzmann constant ($k = 1.38 \times 10^{-23}$ m^2 kg/s^2 T), and T is the temperature in units Kelvin. Note that the only variable for a given energy is the temperature.

At absolute zero, $T = 0$, the term $e^{(E-E_\mathrm{f})/kT}$ is equal to infinity if $E < E_\mathrm{f}$ and zero if $E > E_\mathrm{f}$ (remember that e raised to plus infinity is infinity and e raised to minus infinity is zero). Therefore, the F-D function tells us that at absolute zero all energy levels above E_f are empty, and its probability of occupancy is zero, and all energy levels below E_f are full, with a 100% probability of being occupied, as I said earlier. As the temperature increases to room temperature (300 K), a slight probability exists that there are some electrons in the higher

allowed energy levels, i.e. in the conduction band, and a loss of electrons in the allowed energy levels, the valence band. The F-D function is symmetrical, as it should be. It confirms mathematically what we saw intuitively and graphically in Figure 2.8. You can see right away that the probability of allowed energy bands greater than 0.1 eV above the Fermi level at 300 K is only 2%, and if we go 0.3 eV above the Fermi level, it goes down to about 10^{-7} or just one electron in 10 million sites (Figure 2.13).

Let me now show you graphically how the F-D statistics represent mathematically exactly what I have just explained. Figure 2.14 is the same as Figure 2.13, except that I have exchanged the axes and drawn only one curve: the 300 K (room temperature) curve. Figure 2.14 shows that at the Fermi level, the probability of electrons occupying an energy level is 50%, assuming, of course, that there are allowed energy levels in that range. Allowed energy levels higher than the Fermi level have a lower and lower probability of being occupied (Figure 2.15). As the energy increases, the probability of having an electron occupying a valid energy level decreases, quickly going to zero. If the energy decreases, the opposite is true: at some point, 100% of all the allowed energies will be occupied by electrons. This F-D curve at 300 K never changes. To change the shape, we need to change the temperature, as shown in Figure 2.13. In Figure 2.15, notice that the F-D functions on the right for each energy band are identical to each other and also identical to the one in Figure 2.14, except that they are smaller to fit the figure. In the case of the insulator (A), the probability that there are any electrons in the conduction band is as close to nothing as it can get, if you consider that the energy gap is 3 eV, the Fermi level is in the middle, and the difference between the lowest level and the Fermi level is 1.5 eV. If I insert this energy difference in Eq. (2.1), I get the probability that the lowest energy in the conduction band is occupied:

$$F(\mathrm{E}) = \frac{1}{\frac{1.5\times1.6\times10^{-6}}{e^{1.38\times10^{-23}\times300}}} = 6.7\times10^{-26} \tag{2.2}$$

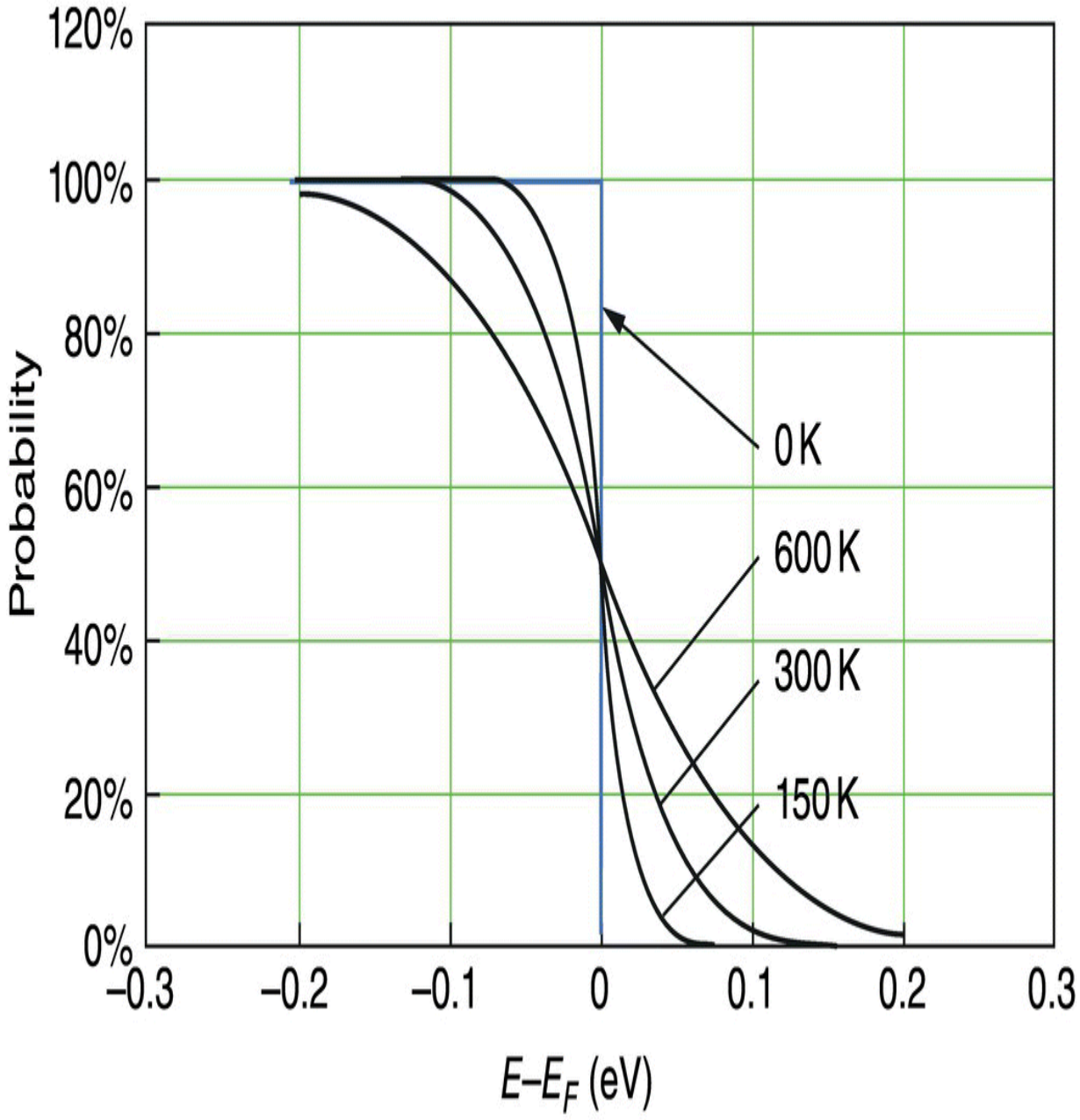


Figure 2.13 The probability that electrons are free as a function of the difference between their energy and the Fermi energy. As the temperature increases, the probability that there are free particle also increases.

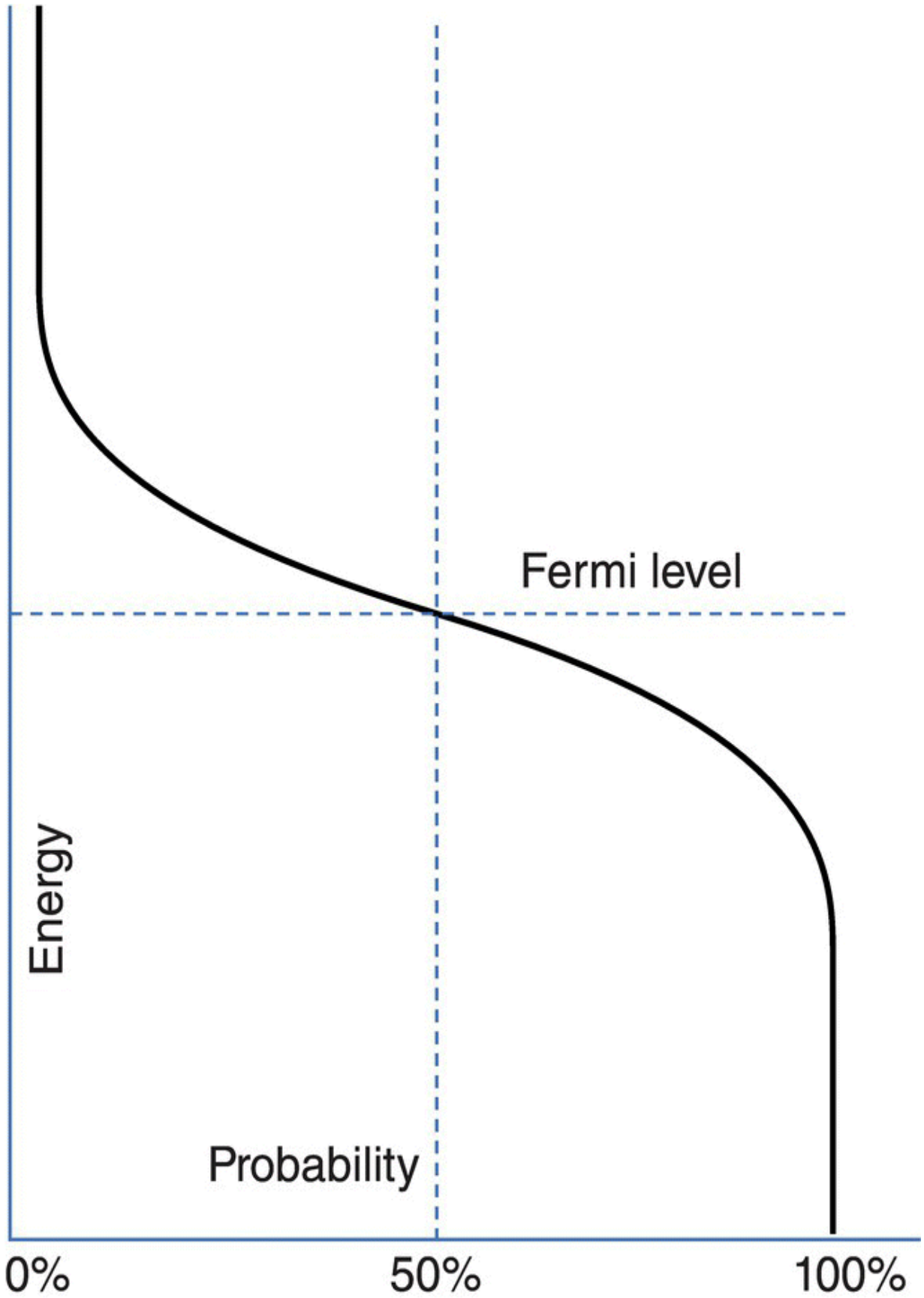

Fermi level
Energy
Probability
0%
50%
100%

Figure 2.14 The F-D function at room temperature.

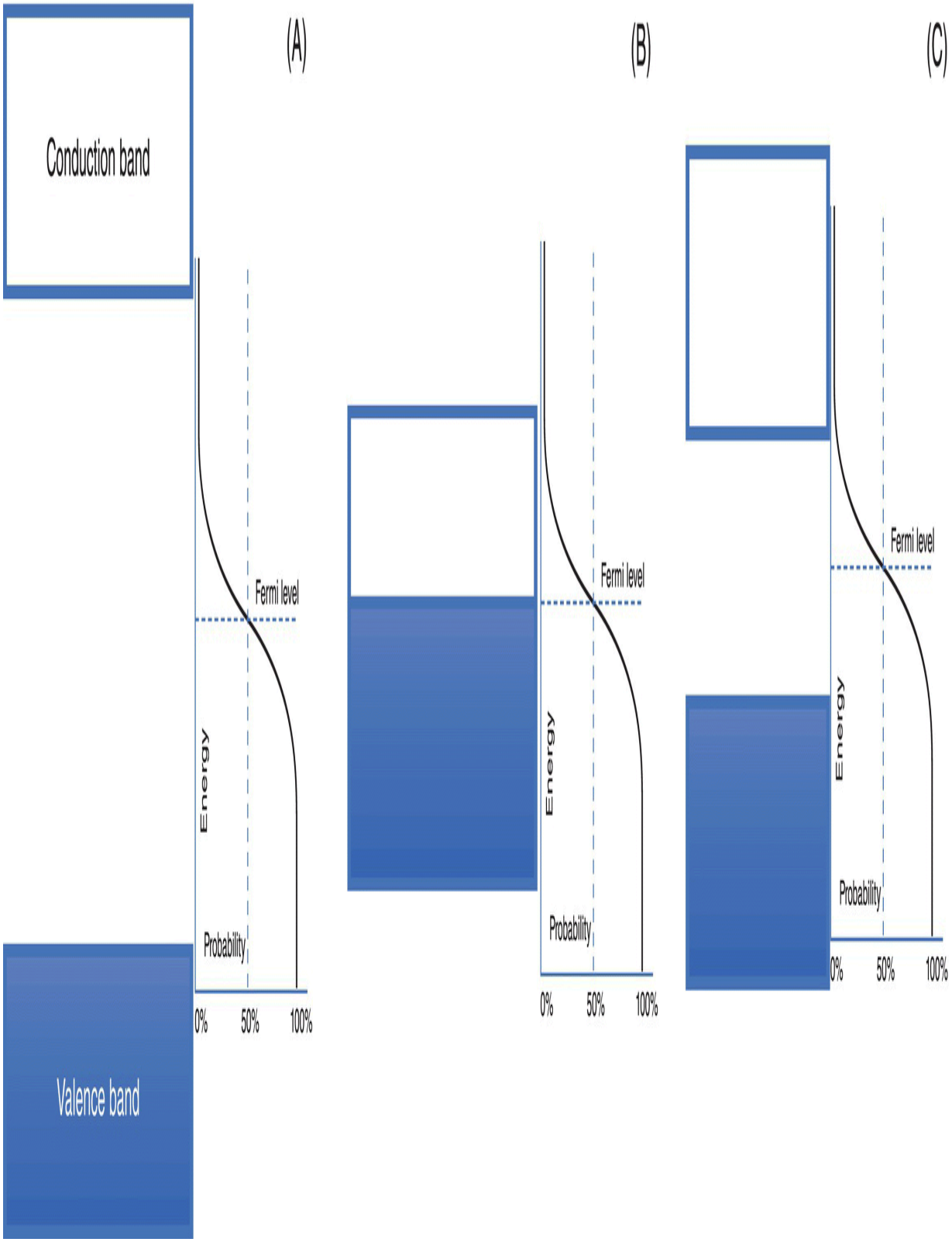


Figure 2.15 The F-D functions on the side of the energy bands of insulators (A), conductors (B), and semiconductors (C) show how many electrons and holes will be at any of the energy values.

This is practically nothing at all.

The case of the conductor (B) is exactly the opposite. At room temperature, the lower energies of a conductor's conduction band are full of electrons, and the valence band has lots of empty sites, that is, holes.

The semiconductor (C) is in the middle. Very few electrons have energies large enough to be in the conduction band. Silicon has an energy gap equal to 1.11 eV. So, since the Fermi level is in the middle, the lowest energy of the silicon's conduction band is 0.555 eV above the Fermi level. If I use this number in Eq. (2.1), the probability that there is an electron in this lowest of the allowed energy location is 5×10^{-10}: very low, but not zero, as I mentioned in Section 2.4.

The same statistical analysis tells us that the product of electrons and holes is given by

$$np = Ce^{-Eg/kT} \tag{2.3}$$

where *C* is a constant, *Eg* is the energy difference between the conduction and valence bands (that is, the energy gap), *k* is the Boltzmann constant, and *T* is the temperature in degrees Kelvin. Notice that *C*, *Eg*, and *k* are all constants. The only variable in this equation is the temperature, *T*. In an intrinsic semiconductor – that is, one without any impurities whatsoever – the number of electrons must be identical to the number of holes:

$$n_i = p_i \tag{2.4}$$

Therefore,

$$n_i^2 = np = Ce^{-Eg/kT} \tag{2.5}$$

that is, the product of the number of free electrons and free holes in a material is always the same at a given temperature.

The number of intrinsic charges in a semiconductor changes drastically with the temperature because the temperature appears in the denominator of the exponential function. Figure 2.9 shows how much the number of free charges in a pure, intrinsic semiconductor like silicon or GaAs changes with temperature. Note that the *y* axis is logarithmic.

3
Types of Semiconductors

OBJECTIVES OF THIS CHAPTER

In the previous chapter we saw that the energy levels in gaseous atoms become energy bands when they form a solid and atoms and electrons interact with each other. The band concept allowed us to understand the difference between conductors, insulators, and semiconductors depending on how full or empty the allowed energy bands are and how large or small the gap between conduction and valence band is.

In this chapter we deal with semiconductors and how I can change their electrical properties by purposely adding impurities. One of the great advantages of the semiconductors is that I can dope, that is, add a control number of impurities, and change drastically the electrical properties of the semiconductors. We will look at the crystallographic formation and how electron hold bonds with neighboring atoms.

3.1 Semiconductor Materials

There are many semiconductor materials, each having its own advantages and disadvantages. Figure 1.17 I showed a portion of the Mendeleev's periodic table that includes all the elements that we use as semiconductors. The single element semiconductors are those in group IV. Group IV elements are those materials that have four valence electrons, that is, four electrons in the outermost orbit or on the valence band. These are also the main semiconductor materials we use, silicon and germanium. There are also compound

semiconductors where the number of valence electrons averages four (e.g. groups III and V since [3 + 5]/2 = 4). Such compounds are formed with elements from groups III and V, or groups II and VI. The most important of these compound semiconductors is gallium arsenide, GaAs.

I list below only the most commonly used semiconductors.

Element semiconductors are those in group IV of the periodic table. They are:

1.1) *Silicon (Si)*, which has an energy gap of 1.12 eV. Silicon is by far the most commonly used material in microelectronics.

1.2) *Germanium (Ge)*, which has a smaller energy gap, 0.67 eV, than silicon. It was very much used at the beginning of the electronic era.

Compound semiconductors consists of two atoms, one in group III and the other in group V, or one in group II and the other in group VI, so that the compound material has an average of four electrons in the valence band.

2.1) *Gallium arsenide (GaAs)* is the second most used semiconductor material. The energy gap is larger than silicon, 1.43 eV, which is very useful for some applications such as microwave, lasers, and some types of very efficient solar cells. It is difficult and expensive to grow very pure GaAs material.

2.2) *Indium antimonite (InSb)* has a very narrow energy gap, only 0.17 eV. It is widely used in infrared detectors.

2.3) Less common are *indium arsenide (InAs)*, with an energy gap of 0.36 eV, and *cadmium telluride (CaTe)*, with energy gap of 1.49 eV, which is used in solar cells.

There are also some tertiary compounds, like *HgCaTe* (mercury–cadmium–tellurite). It has the great advantage that its energy gap changes depending on the composition, from almost 0 to 1.5 eV. It is almost exclusively and widely used in infrared detectors. I discuss HgCaTe detectors in Section 4.6.

What is common for all these semiconductors is that they have either four valence electrons or, in compound semiconductors, their average number of electrons in the valence band is four.

3.2 Short Summary of Semiconductor Materials

3.2.1 Silicon

By far the most commonly used semiconductor material is silicon. I will use silicon to explain semiconductor technology throughout this book.

Silicon has many advantages:

Silicon is the second most common element on earth (28%) after oxygen (46%).

The many years of use has made silicon the cheapest and the most used material in electronics.

Silicon can be grown to be very pure.

Silicon wafers can be very large, 300 mm (almost a foot), with research now being done to grow them to 450 mm.

Silicon is hard and strong so we can handle the wafers easily (with lots of care, of course).

Silicon has a natural oxide that is easy to grow, matches the structure of silicon, and has good insolating and electric properties.

We can etch the oxide to make well-defined openings in it. This allows us to change locally the properties of silicon by inserting a variety of desired impurities.

I cover many of these properties in Chapter 10 when I discuss the fabrication of integrated circuits. Most of what I say about silicon is equally relevant for the other semiconductors. Figure 3.1 shows the crystallographic structure of silicon.

The diamond crystal structure of silicon is one of the strongest. If you look at the black atoms inside the dashed box in Figure 3.1, you will see that the atom at the center is joined to the four closest neighbor atoms at each corner of the dashed box. Each atom, as we have seen, has four valence electrons, so they grab each other and form what we call a *covalent* bond. By sharing electrons, each atom completes the outer shell in a very strong and stable configuration.

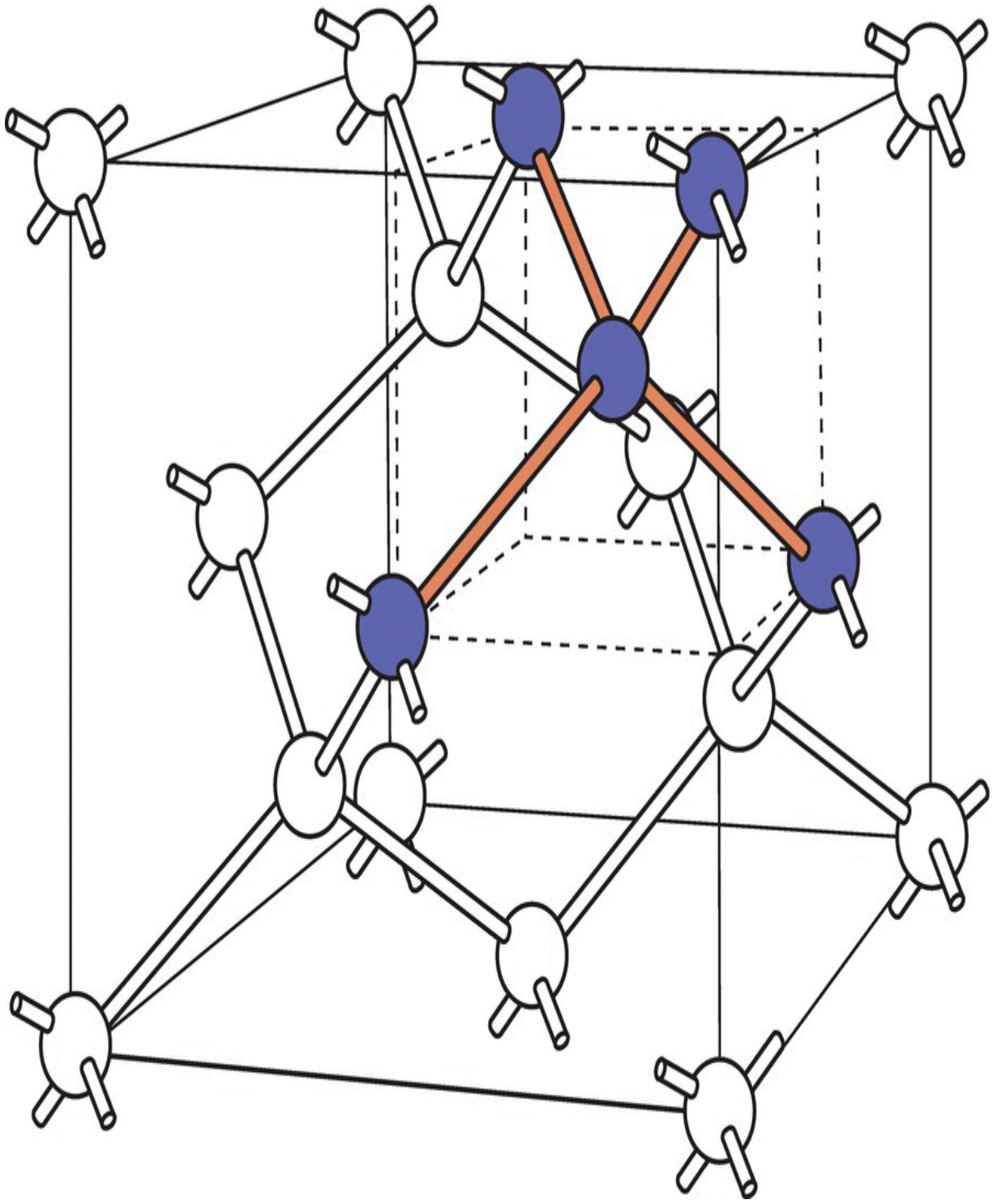


Figure 3.1 Diamond crystal structure of silicon and germanium. The black balls show one single atom and the use of its four valence electrons (solid lines) to bond with the surrounding atoms.

3.2.2 Germanium

Germanium was discovered in 1886, a mere 135 years ago, by Clemens Winkler (1838–1904; Figure 3.2) a German chemist, who called his discovered element germanium (he was obviously German).

It is interesting that although now the element Ge is very well known, there were more than 70 other elements known before Winkler discovered Ge, 18 years after Mendeleev published his periodic table in 1864, Figure 1.5. Mendeleev left a place-holder in his periodic table for Ge, predicting its existence and basic properties like mass and number of electrons. Germanium has a total of 32 electrons, including the four valence electrons. Its crystallographic structure is also a diamond structure (Figure 3.1), the same as silicon.

It is also interesting that the first transistor invented at Bell Labs by a team of engineers led by John Bardeen, William Shockley, and Walter Brattain (Figure 3.3) used Ge. These three American scientists won the Nobel prize for developing transistor action in 1956. They took two gold foils and attached them to two sides of a triangular isolator so that there was little distance between the two foils at the bottom of the triangle. They set that over a piece of germanium and demonstrated that the current between the foils could be modulated by the voltage in the germanium, that is, they demonstrated amplification.

The silicon material is so abundant and inexpensive, with better thermal properties to get rid of heat, that other materials are only used for minor and specialized applications. Now that material growing techniques have improved enormously, there is more interest in germanium. The electrons in Ge move about 10 times faster than in silicon and the holes move four times faster. Now that higher computer speed are being demanded, this germanium property makes germanium devices more desirable.

Figure 3.2 Clemens Winkler, who discovered the element germanium.

Source: Wikipedia, https://en.wikipedia.org/wiki/Clemens_Winkler#/media/File:Winkler_Clemens.jpg.

Figure 3.3 John Bardeen, William Shockley, and Walter Brattain at Bell labs in 1948 (left) and a replica of their first transistor (right).

Source: https://en.wikipedia.org/wiki/John_Bardeen#/media/File:Bardeen_Shockley_Brattain_1948.JPG (left); https://upload.wikimedia.org/wikipedia/commons/b/bf/Replica-of-first-transistor.jpg (right).

Because both silicon and germanium have the same crystallographic structure it is possible to grow germanium on top of a silicon wafer. The lattice constant of silicon is 5.43 Å and that of germanium is 5.66 Å. Not the same but close to each other.

Germanium is also used in infrared detectors and spectroscopes.

3.2.3 Gallium Arsenide

GaAs compared to Si is the opposite of Ge. Its energy gap is larger than that of Si, which makes it suitable for many applications, including microwave and laser diodes. It can absorb light photons and emit them more easily than Si can. The crystallographic structure of GaAs (Figure 3.4) is called zincblende and it is very similar to the diamond structure of Si. The only difference is that a row of gallium interpenetrates a row of arsenic but again, each atom is attached to the surrounding ones with by four valence electrons.

The bonding is due to the fact that the arsenic gives up one electron to the gallium so that both gallium and arsenic share the required four electrons to form similar covalent bonding to that in silicon and germanium. The transfer of one or more electrons from one atom to another to bond them together is called *ionic bonding*. Most of the group III–V semiconductor compounds (InSb, GaAs, GaSb, InP, GaP, etc.) have this crystallographic structure.

The planar drawing of Figure 3.5 shows how the electrons form a bond in a II–VI compound. Two neighboring tellurium atoms lend one electron each to the cadmium atom to form the strong zincblende crystal structure.

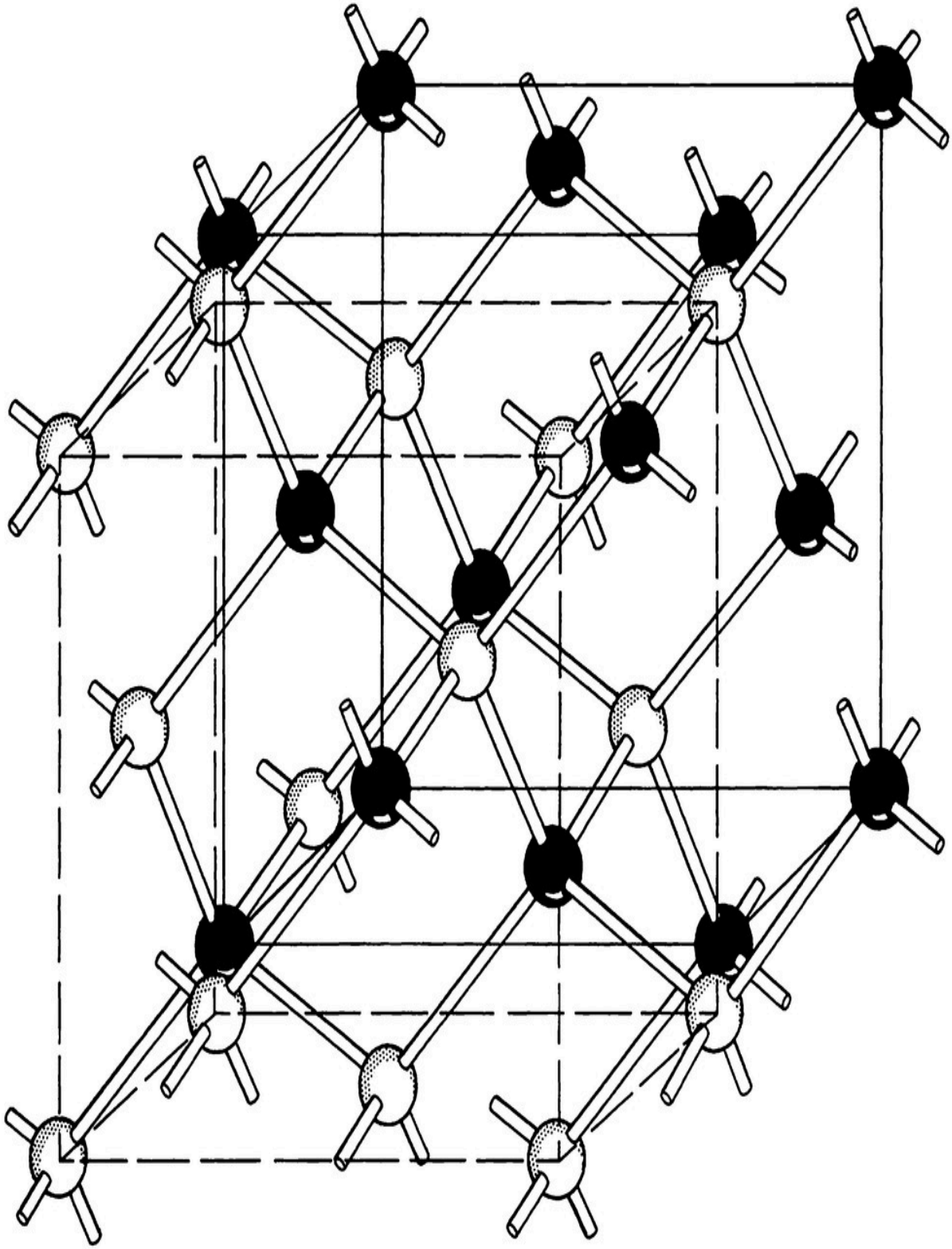


Figure 3.4 The zincblende structure of GaAs is very similar to that of Si, the diamond crystal structure.

3.3 Intrinsic Semiconductors

By an intrinsic semiconductor we mean a semiconductor crystal composed of just one set of atoms, Si atoms or GaAs atoms, for example, without any impurities or crystallographic defects whatsoever.

Silicon has four valence electrons. This makes the crystal structure (Figure 3.1) one of the strongest bonds between atoms, the same structure as diamond. The four electrons are in shell number 3p, a shell that has space for eight electrons. These four are the valence electrons. As we grow a pure crystal, silicon atoms naturally bond with the surrounding atoms so that each atom shares eight electrons with its four neighbors. This sharing results in a complete shell. Figure 3.1 shows graphically the three-dimensional arrangement of atoms in an Si crystal. The blackened lines and balls show one atom, with the valence electrons bonding firmly with the surrounding four atoms. Figure 3.6 is a more abstract two-dimensional representation of an Si atom with the four valence electrons represented as lines around the nucleus.

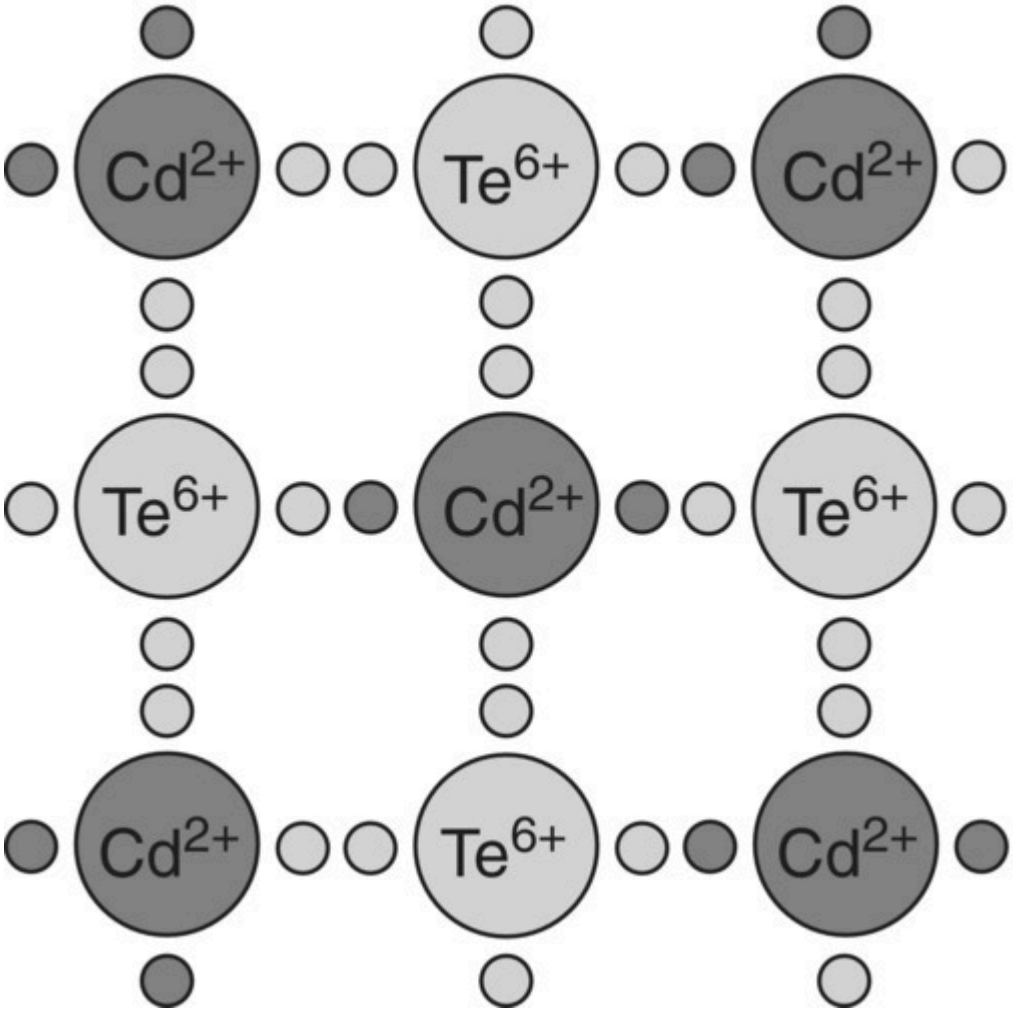


Figure 3.5 The unit structure of CdTe shows how the cadmium, valence two, and the tellurium, valence six, combine in a solid. The two-dimensional sketch shows how tellurium generously gives two electrons to the cadmium atom to complete the bonding.

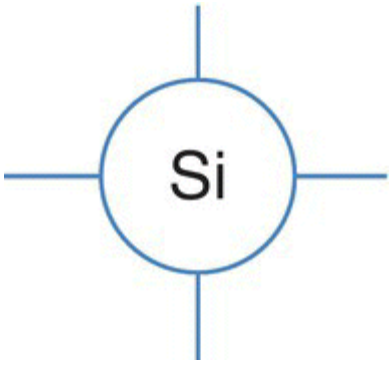


Figure 3.6 The silicon atom has four electrons in the outer shell, shells 3s and 3p, which I represent as lines.

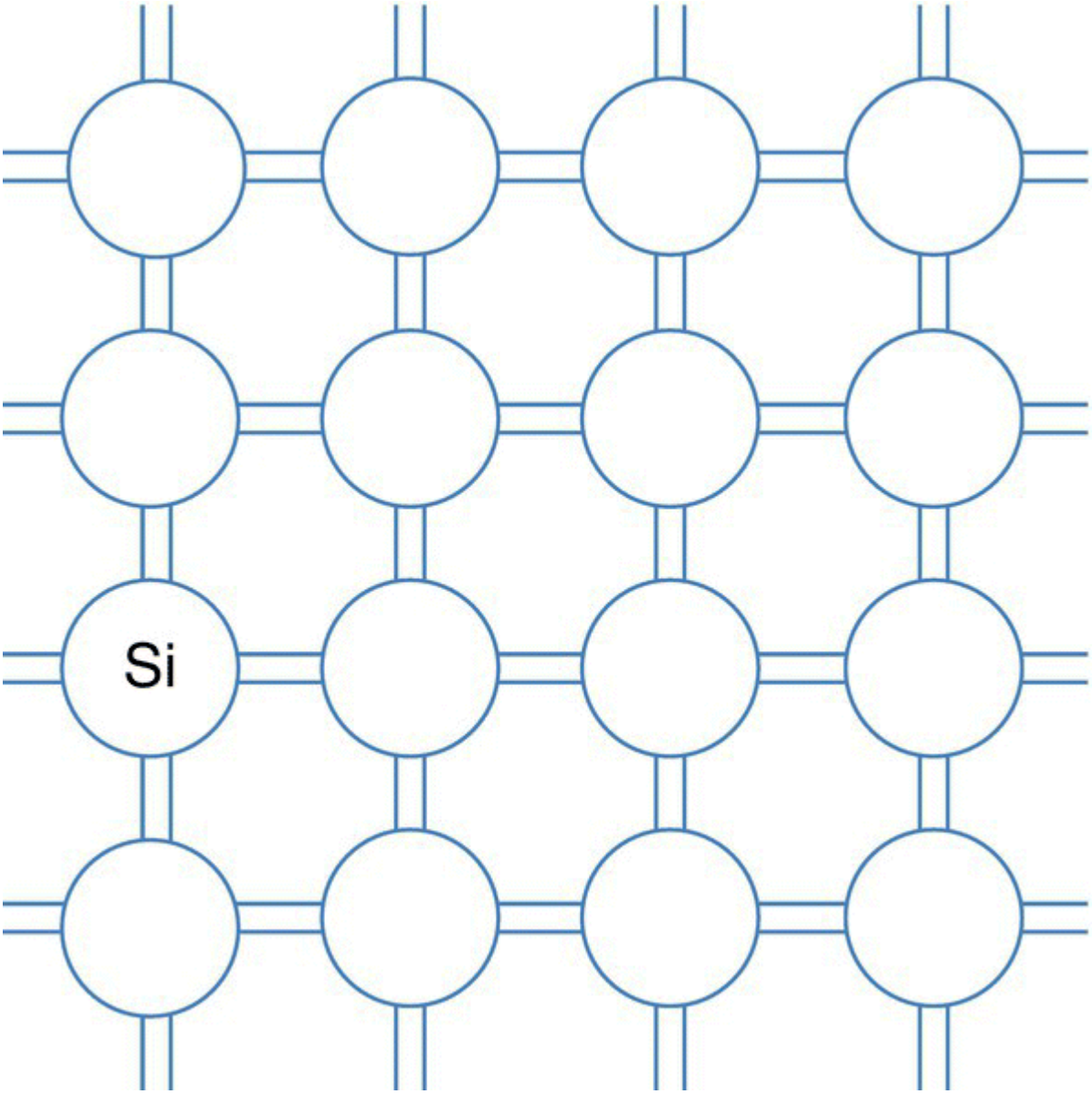


Figure 3.7 A two-dimensional representation of the silicon crystal showing how the electrons form the covalent bonding, completing the 3s and 3p orbits of all the silicon atoms.

Now, using our two-dimensional representation (Figure 3.7), I show how the Si atoms arrange themselves in a crystal.

Each silicon atom shares its four valence electrons with those of the surrounding atoms to form a complete shell. As I said before, this is

called *covalent bonding*. The strong covalent bonding implies that we need a lot of energy to break one of these bonds and get a "free" electron, an electron that can move freely inside the material. This is precisely the concept of the energy gap (*Eg*). The energy gap is a visual representation of the energy that an electron in the valence band needs to break a bond and be free to move if I apply a voltage. A Si electron attached to its atom needs 1.12 eV of energy to break the bond. As I said before, at room temperature, statistically, only 1.45×10^{10} electrons per cm^3 have sufficient energy to break this bond.

By adding a controlled amount of impurities to the silicon, or other semiconductor materials, I can change drastically the electrical properties of the semiconductor, as I explain in the next section.

3.4 Doped Semiconductors: n-Type

The elements in group V in the periodic table (nitrogen, phosphorous, arsenic, and antimony) have five electrons in the valence band. Suppose that, as I grow a Si crystal, I add a small, but controlled, amount of antimony (Sb) atoms, and I mean a very tiny amount, for example 1×10^{16} Sb atoms per cm^3, still much larger than the number of intrinsic electrons, $n_{i,}$ but much smaller than the total number of Si atoms. Sb has five electrons in the outer orbit ready to bond with whatever other atoms are present. Note that I am talking about one Sb atom for every 5×10^8 (500 million) Si atoms. The lonely Sb atom is completely surrounded by Si atoms. It has no other choice but to replace a Si atom and bond with the surrounding atoms; if you can't beat them, join them. It uses four of its five valence electrons to bond with the four valence electrons of the Si, but what happens to the fifth electron? See Figure 3.8.

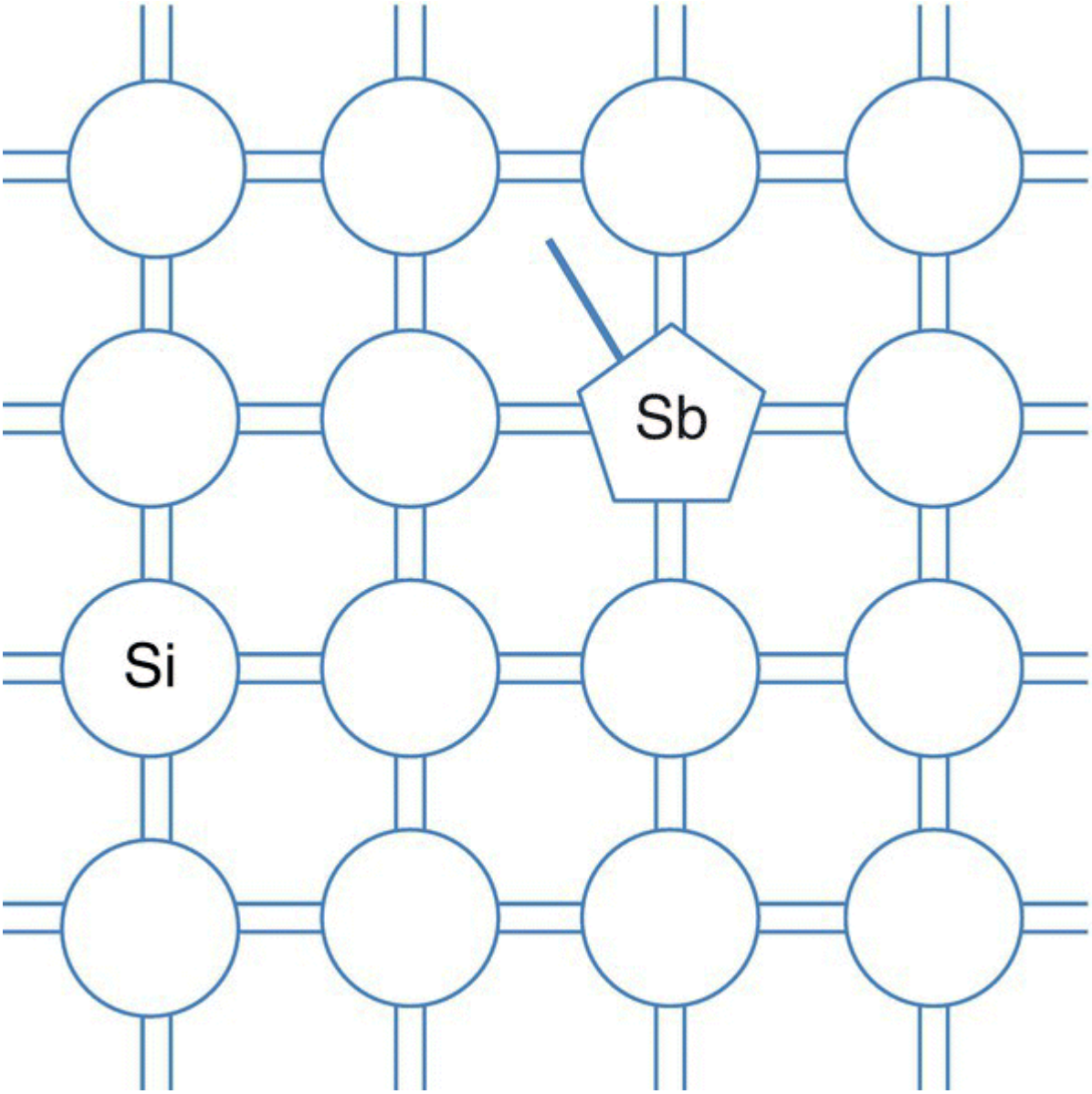


Figure 3.8 A lonely Sb atom in a sea of Si atoms bonds to the surrounding Si atoms with four of its five valence electrons, leaving the fifth electron without a partner.

At absolute zero temperature, the lonely fifth electron, which I show as a bold line sticking out from the Sb atom in Figure 3.8, is attached to its own Sb atom, but this bond is very weak compared to the other four bonds and needs very little energy to break away. What that means is that at room temperature all these fifth electrons have sufficient energy to break the bond and move to the conduction band. Thus, at room temperature there are as many free

electrons in the conduction band as there are Sb atoms in the crystal, one free electron for each of the Sb atoms. If we call the density of Sb atoms N_D (D because these atoms are called *donors*, that is, they donate a free electron to the crystal) then the number of free electrons in the conduction band is very close to the number of Sb atoms, or $n \approx N_D$. (The wiggly equal sign means that the two numbers are almost the same, or almost equal to each other.) Appendix 3.2 explains why the number of free electrons in the conduction band is almost equal to the number of impurity atoms. I show this much smaller ionization energy in Figure 3.9.

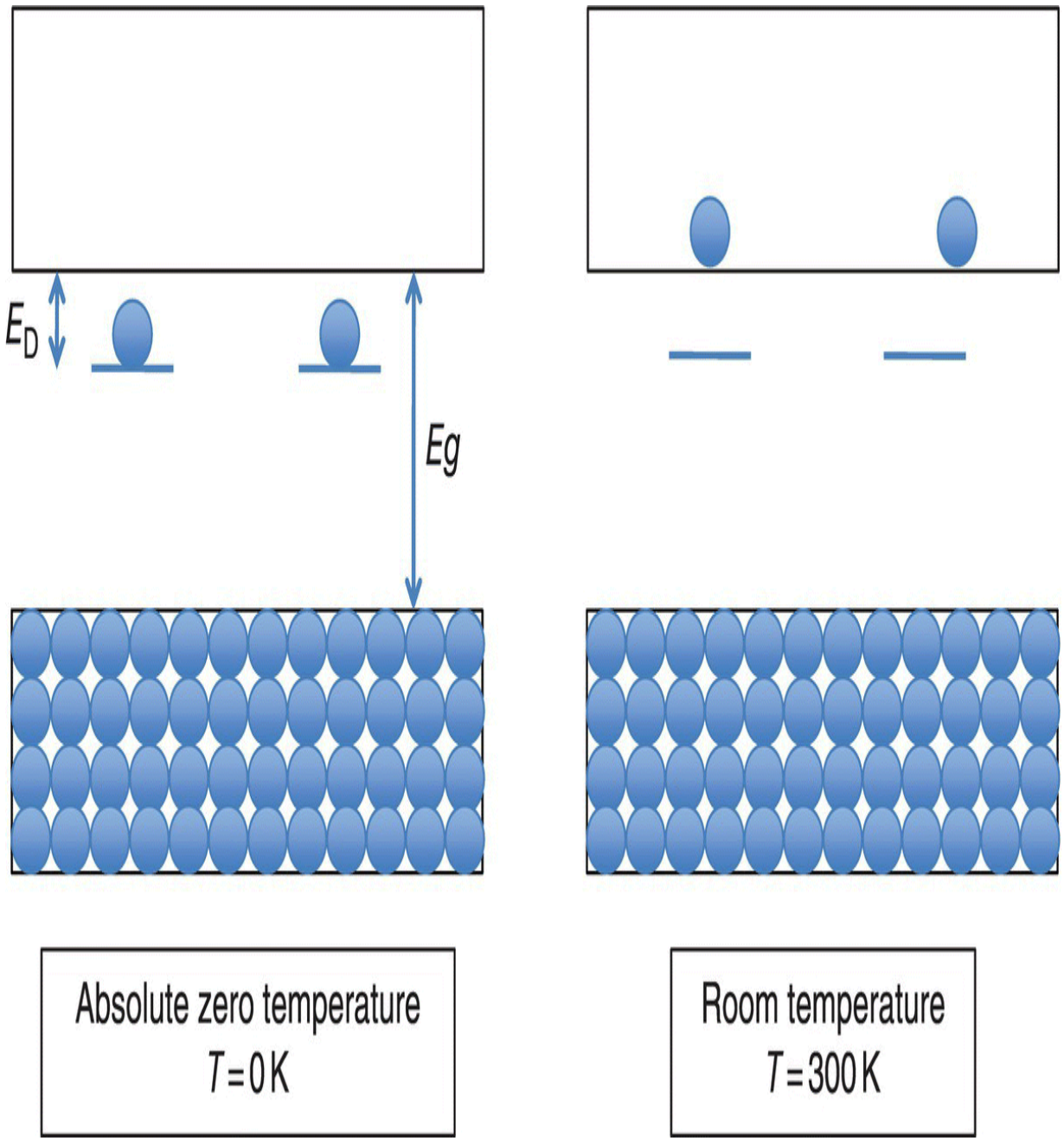


Figure 3.9 Energy diagram of a semiconductor doped with donor atoms. At absolute zero (left) the electrons are bonded to their own atom, but they need very little energy, E_D, to break the bond and at room temperature (right) they move to the conduction band and are free to move.

There are several things I would like to emphasize in the energy diagram of Figure 3.9. First, remember that the whole idea of the

energy gap, *Eg*, is a graphical representation of the energy that an electron in the valence band needs to break the covalent bond and jump from the valence to the conduction band. Now we have to represent the energy that it takes to free the fifth electron from the Sb atom. It is obvious that the energy needed to break the fifth bond is much lower than that needed to kick the electron from the valence band. So we represent this much smaller energy by adding discrete energy levels much closer to the conduction band. We have already seen that the energy needed to free an electron from the silicon is Eg = 1.12 eV. The energy needed to free the fifth electron from the Sb atom is almost 30 times smaller, E_D = 0.039 eV. We call this level E_D because it is the energy needed to free an electron from a donor atom. Second, the Sb atoms are so far separated from one another that they do not interact between themselves. They are like gasses trapped inside a silicon structure. Therefore, I represent that situation with an energy level, not an energy band.

At 0 K there is no energy, so all the electrons are attached to their respective atoms, as I show on the left of Figure 3.9. There are no electrons in the conduction band, but at room temperature, 300 K, there is sufficient energy to ionize, that is, to free all the fifth electrons from the Sb atoms. I show this condition on the right of Figure 3.9. All the energy levels of the fifth electrons are empty. I ignore the tiny number of electrons that come from the valence band. Appendix 3.2 explains more precisely why all the electrons in the donor level go to the conduction band at room temperature.

The number of free electrons, n, in the conduction band is therefore:

$$n = N_D + n_i \approx N_D = 1 \times 10^{16} \tag{3.1}$$

since n_i is insignificant compared to N_D, the number of impurity atoms.

I showed in Appendix 2.2 that the product of the number of electrons, n, times the number of holes, p, is constant. Thus

$$np = n_i p_i = n_i^2 = \left(1.45\times10^{10}\right)^2 = 2.1\times10^{20} \tag{3.2}$$

since $n_i = p_i$. Therefore, if the number of electrons, n, in an n-type semiconductor is 10^{16} (Eq. 3.1) then the number of holes, p, is going to be

$$p = \frac{n_i^2}{n} = \frac{2.1\times10^{20}}{10^{16}} = 2.1\times10^{4} \tag{3.3}$$

Look at the ratios (using only the powers of 10):

number of atoms in silicon = 10^{22} per cm^3

number of impurity donors, N_D = 10^{16} per cm^3

number of free electrons in the conduction band at 300 K, n = 10^{16} per cm^3

number of intrinsic charges, n_i = 10^{10} per cm^3

number of holes, p = 10^4 per cm^3.

What a spread. Also note that the number of electrons coming from the valence band has to be the same as the number of holes left in the valence band, and this number is 10^4. So, I am quite justified in ignoring n_i in Eq. (3.1). That gives you an idea of how much I can change the properties of semiconductors.

I have just described what an n-type semiconductor is, n-type because, at room temperature, I have added lots of free electrons in the conduction band and for all practical purposes eliminated the number of holes in the valence band (continuing with our freeway/garage analogy from Chapter 2, there are so many cars parked on the streets that very few of the cars in the garage need to move to the freeway, leaving very few empty spaces in the garage).

One question you may ask is why I added so few Sb atoms. Why not add something like 10^{20} or 10^{21} to make the number of free electrons much larger and closer to the number of Si atoms? Actually, we do this sometimes and we called it a *degenerate* n-type semiconductor (I will use these degenerate semiconductors to create contacts, see Section 10.5). The problem is that as we add more and more Sb atoms, the Sb atoms start interacting with each other, not all of them, but maybe a group of them. When there are very few Sb atoms, the Sb atoms are so separated that they act like a gas and have a single energy level similar to the Bohr atom. But as we get more and more of them together, they start forming their own energy band that can encroach on the silicon's conduction band. These impurity bands are very close to the conduction band, even touching it. If this is the case, the semiconductor starts acting as a conductor because a very large number of electrons are free to move in the conduction band even at temperatures close to 0 K.

3.5 Doped Semiconductors: p-Type

Now consider a different situation. Instead of Sb, we dope the Si with boron (B). Boron has a valence of three, that is, there are only three electrons in the outer band, ready to bond with whatever other atoms are around. Because, as in the n-type semiconductor above, we add very few B atoms, they are forced to replace a Si atom and bond with the other surrounding Si atoms. I show this schematically in Figure 3.10.

Notice what happens. The three electrons of the B atom, the triangular shape atom in Figure 3.10, bond with three of the four surrounding electrons of the Si atoms but there is one incomplete bond (I show it as a question mark). You can also imagine that it takes very little energy for one of the surrounding electrons to jump and fill this empty spot if there is an applied voltage attracting the electrons to the positive terminal. A generous neighboring Si atom gives one of its electrons so that the boron atom can fill its incomplete bonding. But that means that the generous neighboring

atom has lost an electron and is now him who has an incomplete bond waiting for some other close-by neighbor to give him the missing electron. In our energy band structure, I represent this situation by adding an energy level very close to the valence band (Figure 3.11).

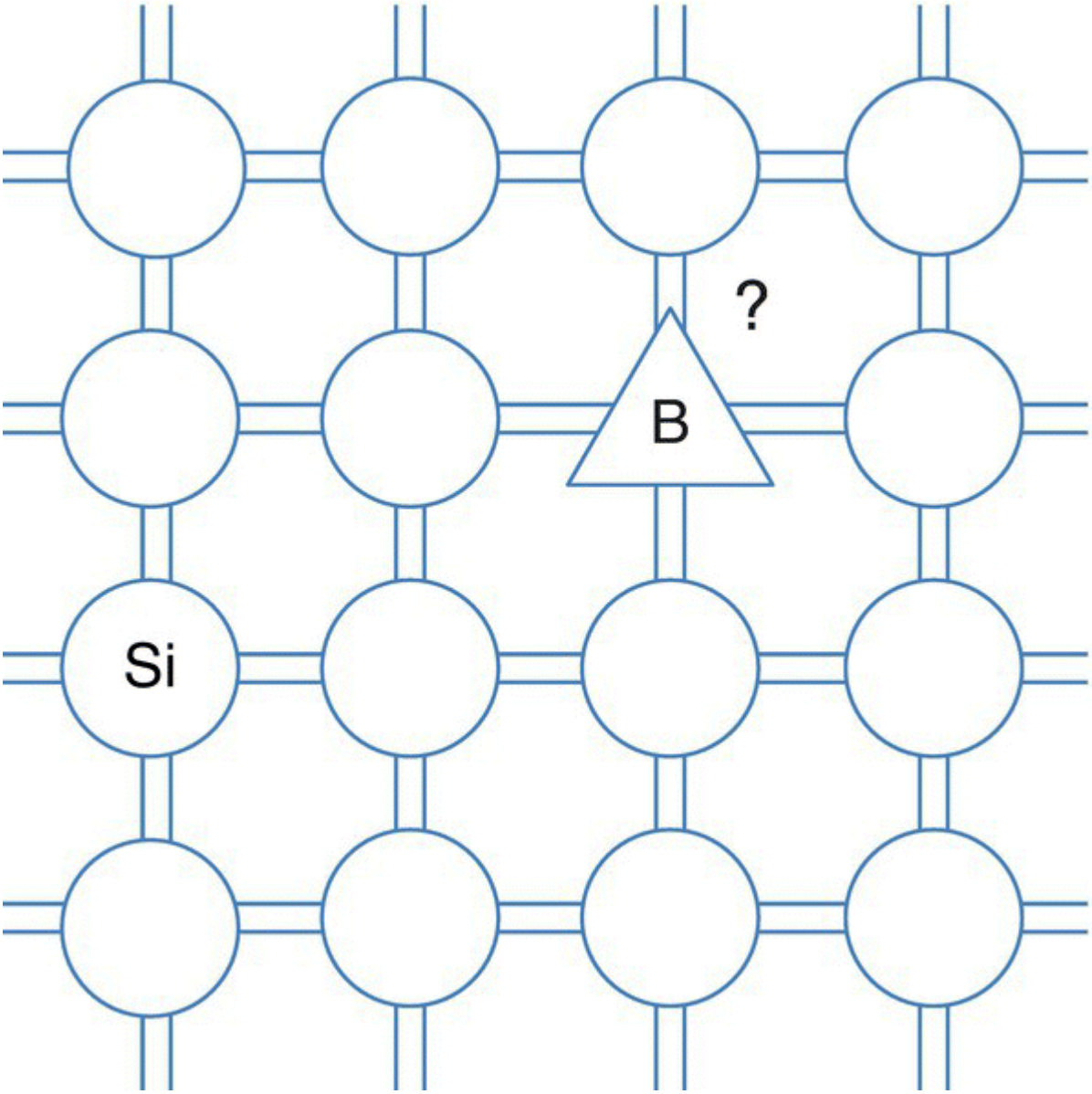


Figure 3.10 The boron atom surrounded by a huge number of Si atoms takes the place of one of them, leaving a missing electron without a bonding partner.

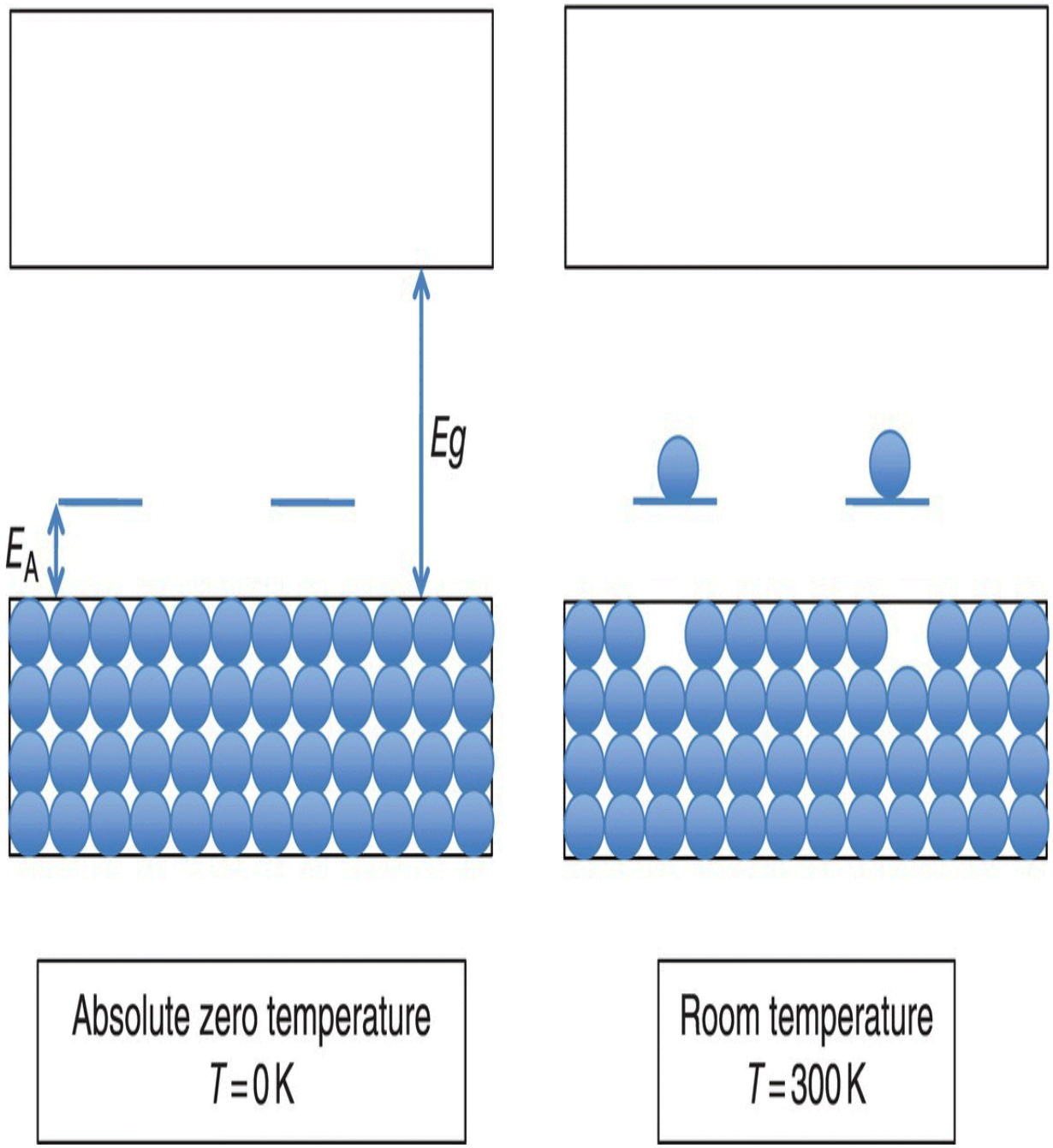


Figure 3.11 The energy of the boron empty bond is very close to the valence band and at room temperature many electrons from the valence band can jump and leave behind empty spaces, holes.

This small energy, E_A (A because these impurities are called acceptor impurities, impurities that accept electrons from surrounding silicon atoms), is the energy that an electron in the valence band (an

electron attached to a silicon atom) needs to jump from one atom to an empty level of a B atom. The left side of Figure 3.11 shows the case when the semiconductor is at 0 K. The electrons are all at their lowest possible energy sites, that is, in the valence band. Nothing moves. But at room temperature, shown on the right of Figure 3.11, many electrons in the valence band have enough energy to jump to the higher, but very close, energy sites created by the boron's empty bonds. The electrons that have moved to the acceptor levels, that is, bonded to the boron atom, cannot move, they are trapped.

Suppose I remove several cars from a full garage to another floor. Now there are empty spaces in the garage and cars can move from one empty space to another without the need to jump to the freeway. The boron impurities at room temperature leave a lot of empty spaces in the valence band and the electrons in the valence band can now move within the valence band. If I apply a positive voltage on the right, the electrons move to the right. The empty, parking, spaces move to the left and therefore it looks from the outside as if positive charges are moving to the left, i.e. to the negative terminal. As I mentioned before (see Section 2.4), we call these empty spaces *holes* and represent them by the letter p. The number of holes is almost identical to the number of indium atoms. Thus $p \approx N_A$, where N_A is the density of acceptor atoms. I show this room temperature condition on the right of Figure 3.11.

It is also important to emphasize that even though we talk about materials that have free electrons or free holes, the material itself is still electrically neutral. Sb has 51 electrons, five of them in the outer, valence, band but it also has 51 protons (positive charges) in the nucleus, therefore the material is electrically neutral (it also happens to have 71 neutrons which have no charge). Similarly, boron has five electrons, three of them in the valence band but it also has five protons and six neutrons in the nucleus, making it a neutral element.

3.6 Additional Considerations

In both the n- and p-type semiconductors I can change the electrical properties by adding impurities. Changing the number of free electrons or free holes also changes the resistivity of the semiconductors. Figure 3.12 shows the resistivity of n- and p-type silicon at room temperature as the number of impurities I add to the silicon increases from 10^{14} to 10^{21} per cm^3. Notice that the scales are logarithmic scales. The resistivity changes by a factor of 1 million, from 0.0004 to 100 Ω-cm. We cannot do that with metals or insulators.

Another point I would like to make is that, in addition to the impurities that we purposely add to the pure semiconductors, there are other impurities which we naturally find in the silicon material and they are quite difficult to get rid of completely. To give you an idea of what we have to deal with, look at Table 3.1.

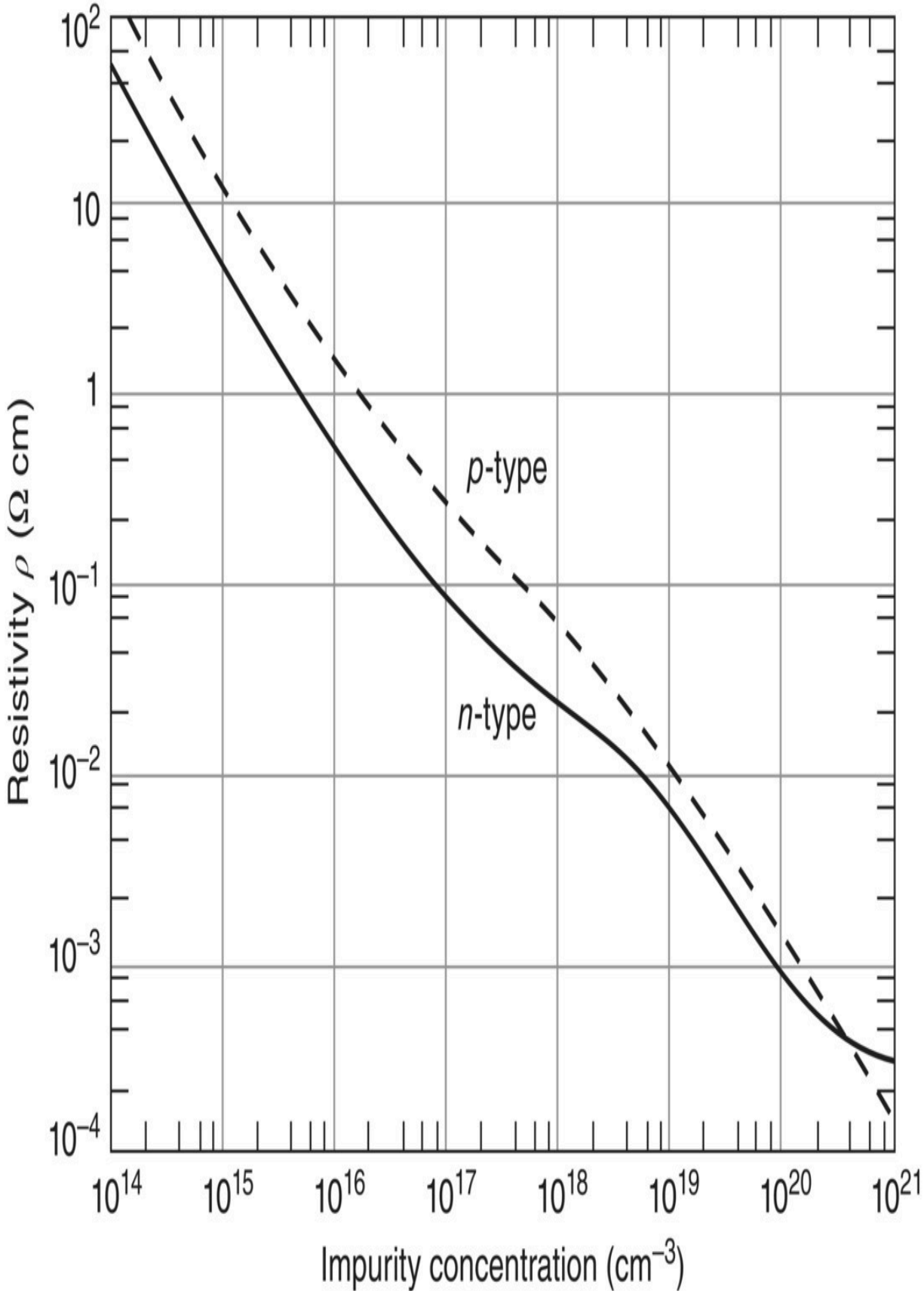

Resistivity ρ (Ω cm)
10^2
10
1
10^-1
10^-2
10^-3
10^-4
p-type
n-type
10^14
10^15
10^16
10^17
10^18
10^19
10^20
10^21
Impurity concentration (cm^-3)

Figure 3.12 the resistivity of n- and p-type silicon changes drastically as the number of impurity atoms increases. As we have seen before, the p-type has a higher resistivity than the n-type.

Table 3.1 The impurities allowed in an electronic grade silicon (parts per billion) is thousands of times better than in metallurgical grade material (parts per million).

Element	Metallurgical grade impurities (ppm)	Electronic grade impurities (ppb)
Boron	50	<0.1
Chromium	200	<0.01
Copper	50	<0.1
Nickle	100	<0.5

First note that the metallurgical grade is measured in ppms, parts per million. Electronic or semiconductor grade silicon is measured in ppbs, parts per billion. By comparison pure water may have arsenic up to 10 ppb, so the electronic purity we need has to be at least 100 times better than the purity of water.

Before we can add a controlled number of impurities, we need to remove those that we do not want. I will talk more about this in Chapter 10 when I discuss the fabrication of integrated circuit (IC) devices. For solar cells or light-emitting diodes, we use an upgraded metallurgical grade, something in between metallurgical and electronic grades, certainly considerably cheaper than using very pure Si. The whole cost of Si is not the extraction of the material, it is ubiquitous, but the purification process.

In addition to impurities that can crawl in or that were there all along and we were not able to completely remove, there are also crystal defects that can be detrimental to semiconductor operation.

One type of defect is the point defects I show in Figure 3.13. Point defects are either *vacancies*, somehow a Si atom is missing, or the opposite, an *interstitial*, that is, an extra atom has squeezed inside

the lattice. As you might expect both types of point defect deform the nicely organized Si lattice. Electron bonding has to be re-established and that may result, like with the added impurities, in effective extra holes or extra electrons. We call these *dangling bonds* because here we have some electrons that do not know who to bond with. Sometimes the two point defects can be related. A silicon atom moves from one location, leaving a vacancy, and moves very close to another location, creating an interstitial.

Another set of imperfections are dislocations, called *staking falls*, as I show in Figure 3.14. On the left of Figure 3.14 there is an additional Si plane that has been inserted and on the right there is an entire plane that is missing. Both are stacking faults. Close to the dislocation the lattice is deformed but further away the lattice is normal.

To complete this topic, let me mention additional impurities that we could use. Figure 3.15 shows graphically and numerically the energy levels of some of these impurities in Si.

Because of its closeness to the conduction or to valence bands, boron, for p-type semiconductors, and antimony and phosphorous, for n-type semiconductors, are the most used dopants in silicon, but we use other impurities for different purposes.

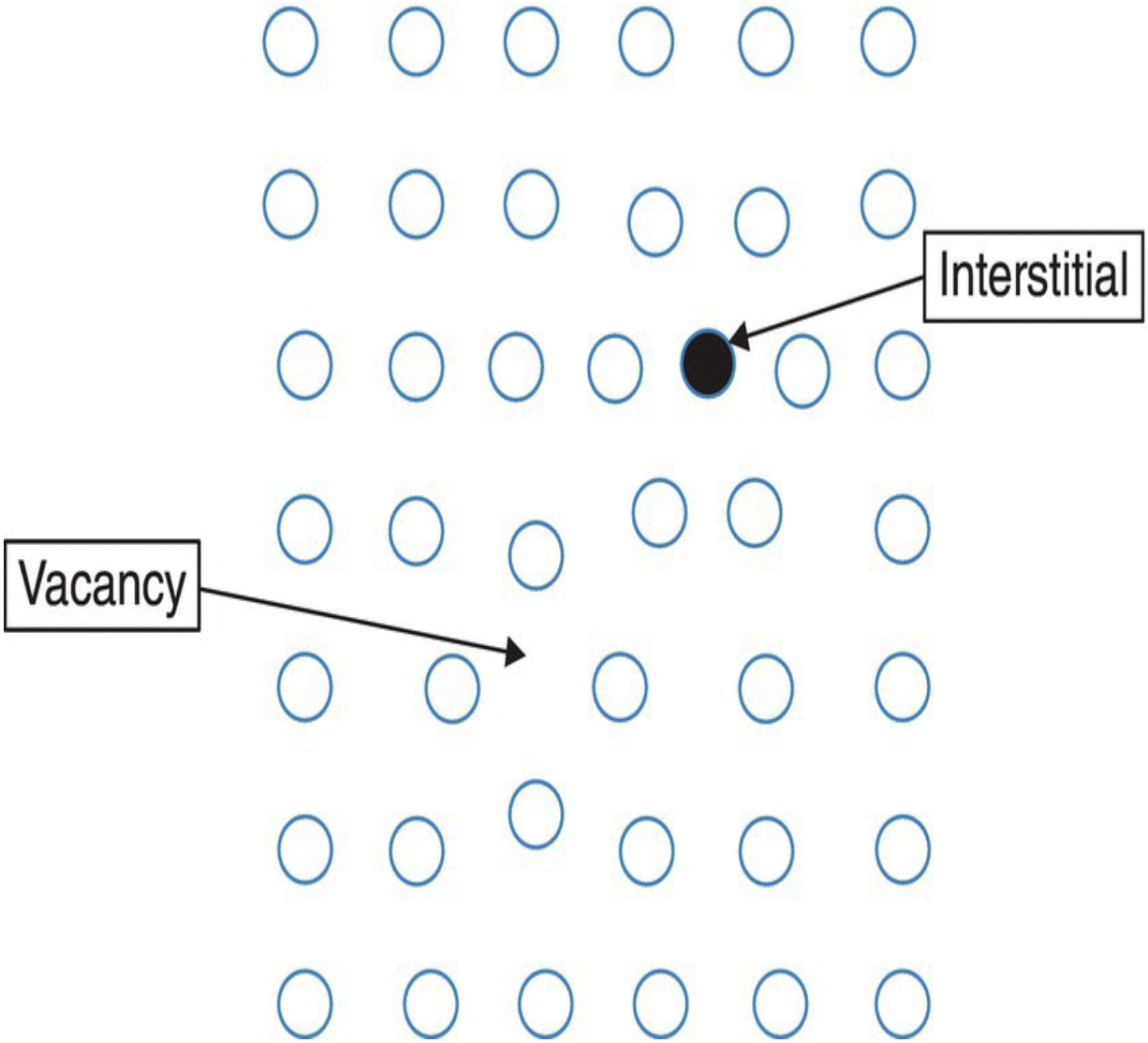


Figure 3.13 Point defects in semiconductors, interstitial atoms or vacancies cause irregularities in the lattice that can act like impurities.

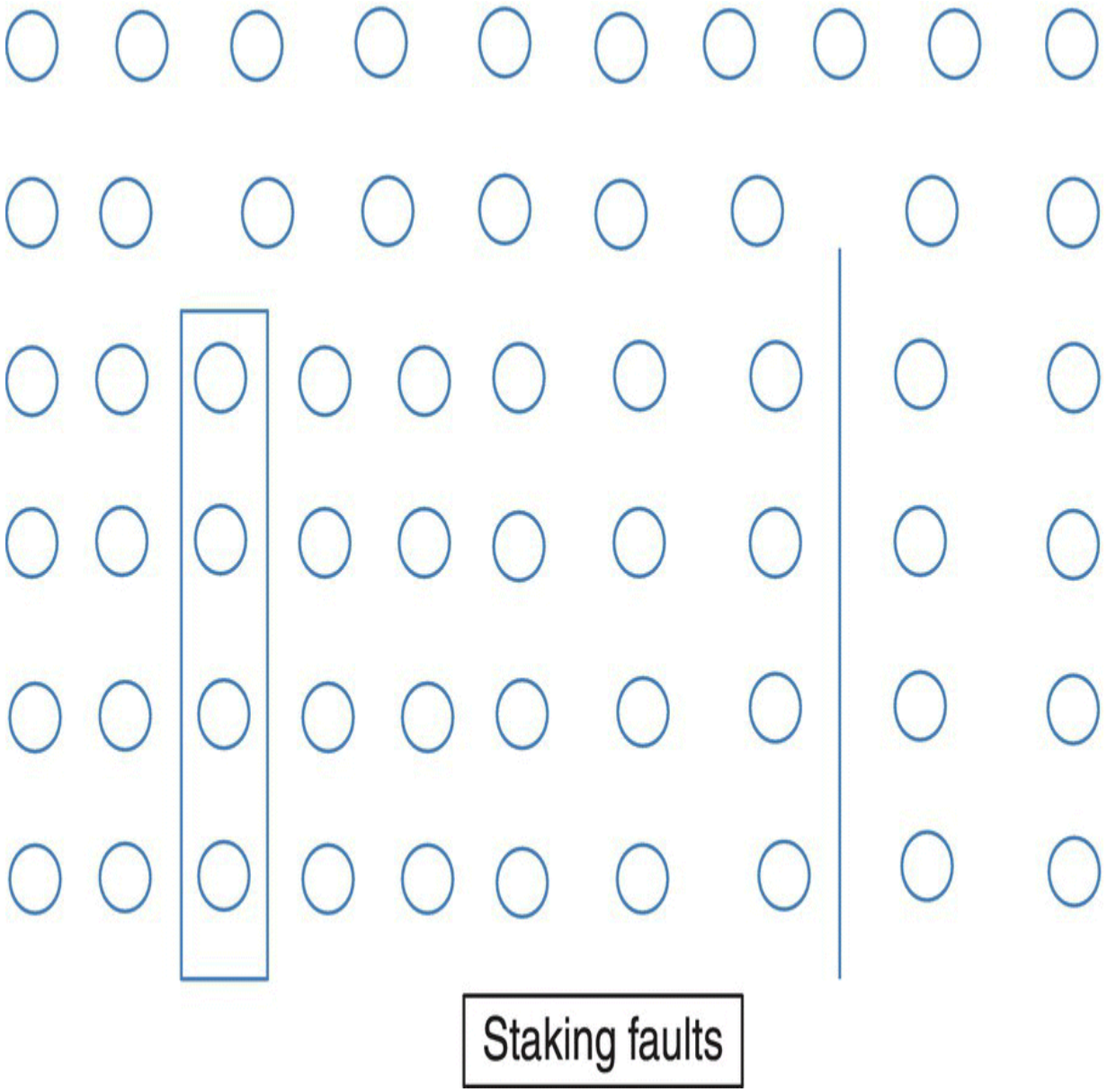


Figure 3.14 Line dislocations, adding or losing a plane of atoms, also cause lattice irregularities that act like impurities.

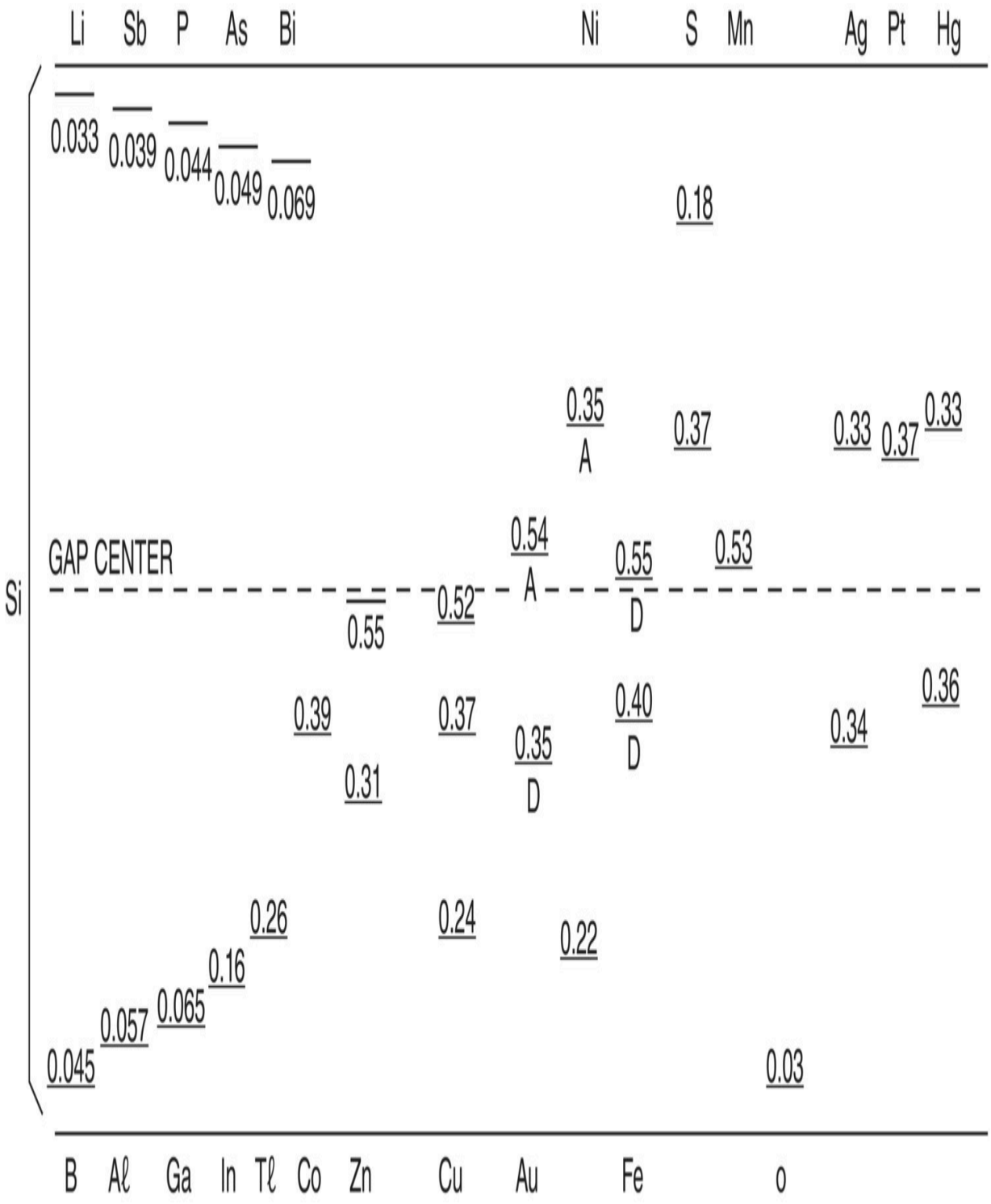


Figure 3.15 There are many native and doped impurities in Si that have very different donor or acceptor energy levels. All the native impurities are undesirable.

3.7 Summary and Conclusions

In this chapter we have seen how I can add a controlled number of impurities to change the semiconductor's characteristics drastically. If I add an atom with five valence electrons in a Si crystal, four of the electrons bond with the Si but there is a fifth electron that needs very little energy to be free. We have also seen that by adding a few controlled numbers of atoms with only three valence electrons we can create a material with a large number of holes. This trick allows us to have semiconductors with a very large range of resistivities. Also, in semiconductor materials we can have electron (negative) or hole (positive) currents.

We will see how we use these properties when we discuss devices: diodes in Chapter 5 and transistors in Chapter 8.

Now let's relax from semiconductor theory. In the next chapter I explain some practical devices that can be understood using the concepts that we have discussed in the first three chapters.

Appendix 3.1 The Fermi Levels in Doped Semiconductors

You should recall from Appendix 2.2 that at absolute zero the Fermi level, E_f, must be halfway between the energy levels occupied by electrons and the energy levels empty of electrons, in pure silicon that is exactly in the middle of the energy gap. In an n-type semiconductor the Fermi level must reside exactly in the middle between the conduction band and the donor impurity energy levels (Figure 3.16) because at 0 K every energy level under the Fermi level must be occupied and every energy level above it must be empty.

The figure on the left of Figure 3.16 is the same as the left of Figure 3.9 and that on the right is the same as the left of Figure 3.11. I have only added in both cases the position of the Fermi level at absolute zero when the electrons are located at their lowest possible energy. On the left I show the case of an n-type semiconductor with

the Fermi level, E_f, between the conduction band, completely empty, and the donor band, completely full. In a p-type material, the Fermi level is between the valence band full of electrons and the empty acceptor levels. If you superimpose the Fermi–Dirac (F-D) function (Figure 2.14) on Figure 3.16 you can see right away that in the n-type semiconductor the number of occupied energy levels in the conduction band at room temperature is very large and the number of empty spaces is very low, and the opposite happens in the p-type semiconductor.

Using the same F-D statistics I discussed in Appendix 2.2, Eq. (2.1), we can show that the concentration of electrons in the conduction band is

$$n = N_D e^{(E_D - E_f)/kT} \tag{3.4}$$

and the concentration of holes in the valence band is

$$p = N_A e^{(E_A - E_f)/kT} \tag{3.5}$$

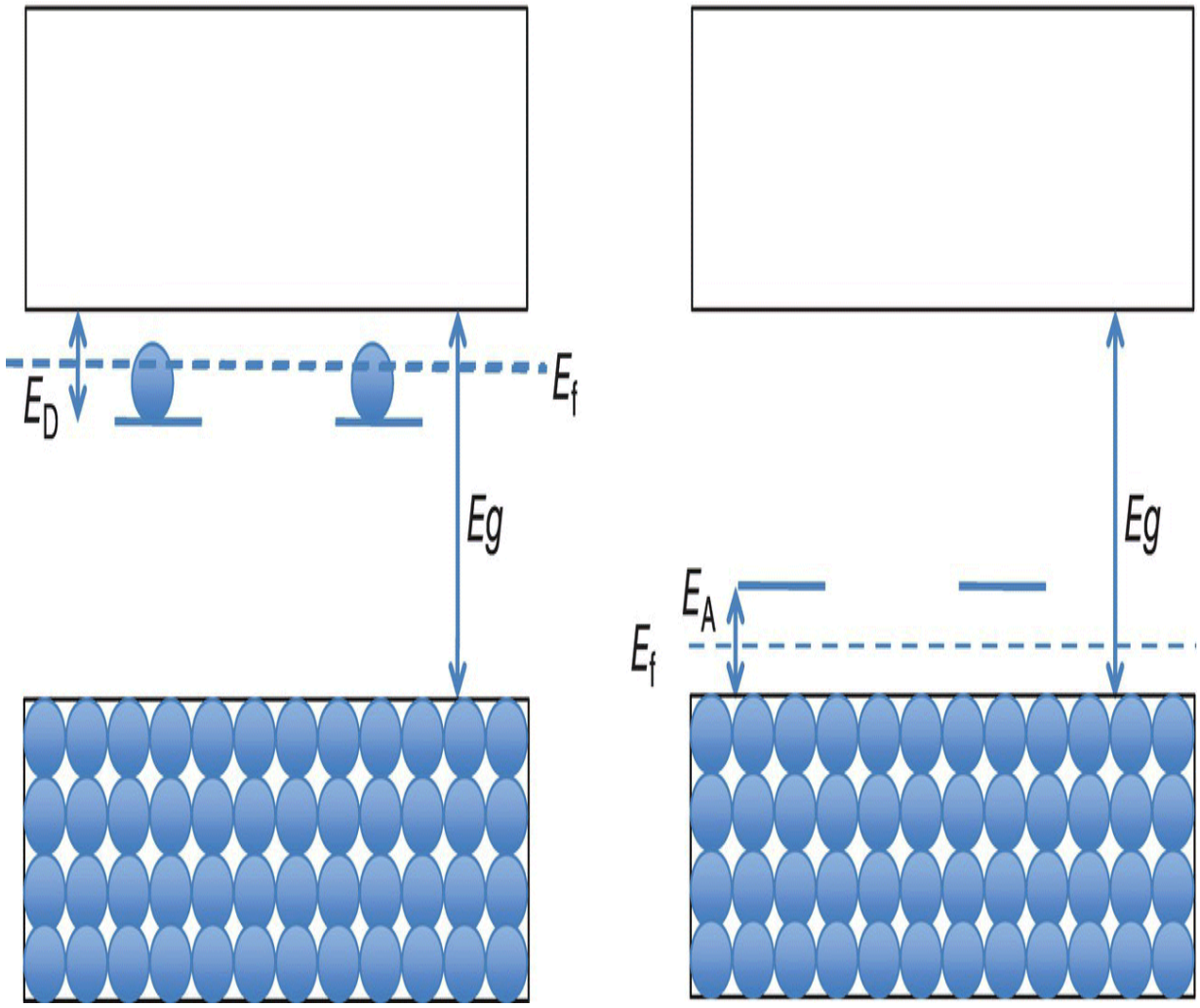


Figure 3.16 The Fermi level in n- and p-type semiconductors at 0 K are in the middle between the donor or acceptor levels and the conduction or valence bands, respectively.

In Appendix 2.2, I showed you how the F-D statistics calculate the number of electrons and holes in insulators, conductors, and semiconductors. Now that we know about doped semiconductors let me show you how the F-D function explains also what is happening in doped semiconductors. Look at Figure 3.17. Figure C is the same as C in Figure 2.15. At 0 K the Fermi level must be between the energy level that has electrons, and the one that does not. Thus, in the case of a donor semiconductor, D, the Fermi level sits between the empty conduction band and the full donor level. The donor level for Sb, for example, is located just 0.039 eV from the conduction

band (see Figure 3.16) so the energy difference between the Fermi level and the conduction band is just 0.02 eV (Figure 3.17D). Let's think about this. If I insert 0.02 eV in Eq. (2.2), I get the result that the probability of occupancy is about 30%. Because the number of donor atoms is much smaller than the number of silicon atoms, 30% means that, for practical purposes, all the electrons in the donor level are now in the conduction band. For holes the situation is different. The distance from the Fermi level to the valence band is 1.09 eV, the energy gap minus half the donor gap (1.11 – 0.02), a huge energy gap and the probability that there are any holes is minuscule (5×10^{-19}) if you do the numbers. The opposite happens with a silicon doped with boron (Figure 3.17E).

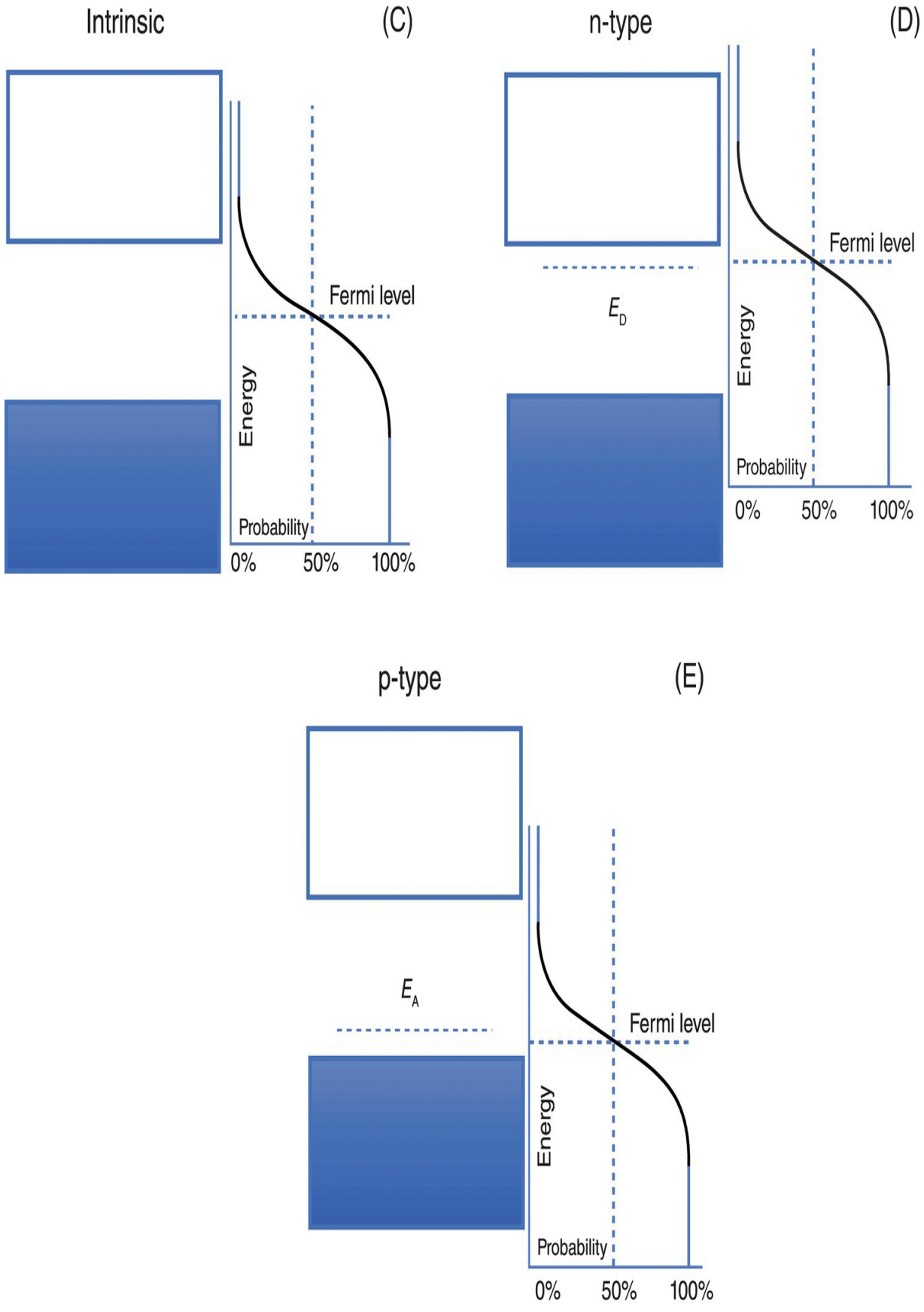

Intrinsic
(C)
Fermi level
Energy
Probability
0%
50%
100%
n-type
(D)
E_D
Fermi level
Energy
Probability
0%
50%
100%
p-type
(E)
E_A
Fermi level
Energy
Probability
0%
50%
100%

Figure 3.17 Intrinsic and doped semiconductors energy bands at 300 K. In the intrinsic semiconductor the Fermi level is located at the exact middle between the conduction and valence bands (C). In the n-type it is between the conduction band and the donor impurity level (D) and in the p-type the fermi level is between the acceptor level and the valence band (E).

Appendix 3.2 Why All Donor Electrons go to the Conduction Band

At this point you may ask why *all* the electrons in the donor levels go to the conduction band. There is not much difference in energy between the conduction band and the donor level. Shouldn't both be almost equally populated with electrons? Why don't just half of them go to the conduction band while the other half remains trapped in the donor levels?

The reason is that I have only 10^{16} donor levels while I have 10^{22} allowed levels in the conduction band, a difference of one million. If a stadium with 300 000 seats is 30% occupied, there will be 90 000 football fans in the stadium, but if we have the same occupation probability, 30%, in a local small theater that has only 65 seats, the number of spectators is only 20.

In a p-type semiconductor the opposite happens. A large number of electrons from the valence band overwhelm the acceptor levels and the number of holes in the valence band is high, limited only by the availability of acceptor levels.

4
Infrared Detectors

OBJECTIVES OF THIS CHAPTER

Let's take a pause from the "semiconductor theory" that I covered in the previous three chapters and discuss one of the applications, infrared detectors, that can be easily understood after we have learned about the concept of energy bands and energy gaps for both intrinsic and extrinsic semiconductors. We do not need to know how semiconductor devices such as transistors or diodes work.

In this chapter, after I explain what infrared radiation is and some infrared applications, I discuss the radiation spectrum and how we use semiconductors, both intrinsic (pure) and extrinsic (doped), to "see" the infrared images. I also cover a couple of compound semiconductors that we use as infrared detectors.

4.1 What is Infrared Radiation?

In 1800 Frederick William Herschel (1738–1822) performed the experiment I show in Figure 4.1. Hershel passed light through a prism that separates the white light into different colors (in a prism, each color bends at slightly different angle, resulting in the color separation, see Appendix 4.1). Hershel placed a thermometer outside the visible radiation on the left of the color red and detected an increase in temperature. Although he did not see anything, it was obvious that there was an "invisible" radiation below (infra) the red color. Very soon scientists realized that the radiation spectrum was considerably wider than just the narrow band of what we call light and that we are able to see.

Figure 4.2 shows the full range of the radiation spectrum, from high frequency gamma rays to very low frequency radio waves. The wavelengths of the entire radiation spectrum, that is, the distance

between two peaks of the waves, λ, range from 10^{-11} m, to 10^{5} m, a whooping 14 orders of magnitude. The visible range is a tiny fraction of the radiation spectrum. The wavelength range of visible light goes from 3.8×10^{-7} m for violet to 7.5×10^{-7} m for red (or from 0.38 to 0.75 μm). Figure 4.2 expands both the visible and the infrared radiation bands. As you can see, infrared radiation has a considerable larger range than visible light, from 7.5×10^{-7} m to 10^{-4} m (or from 0.75 to 100 μm).

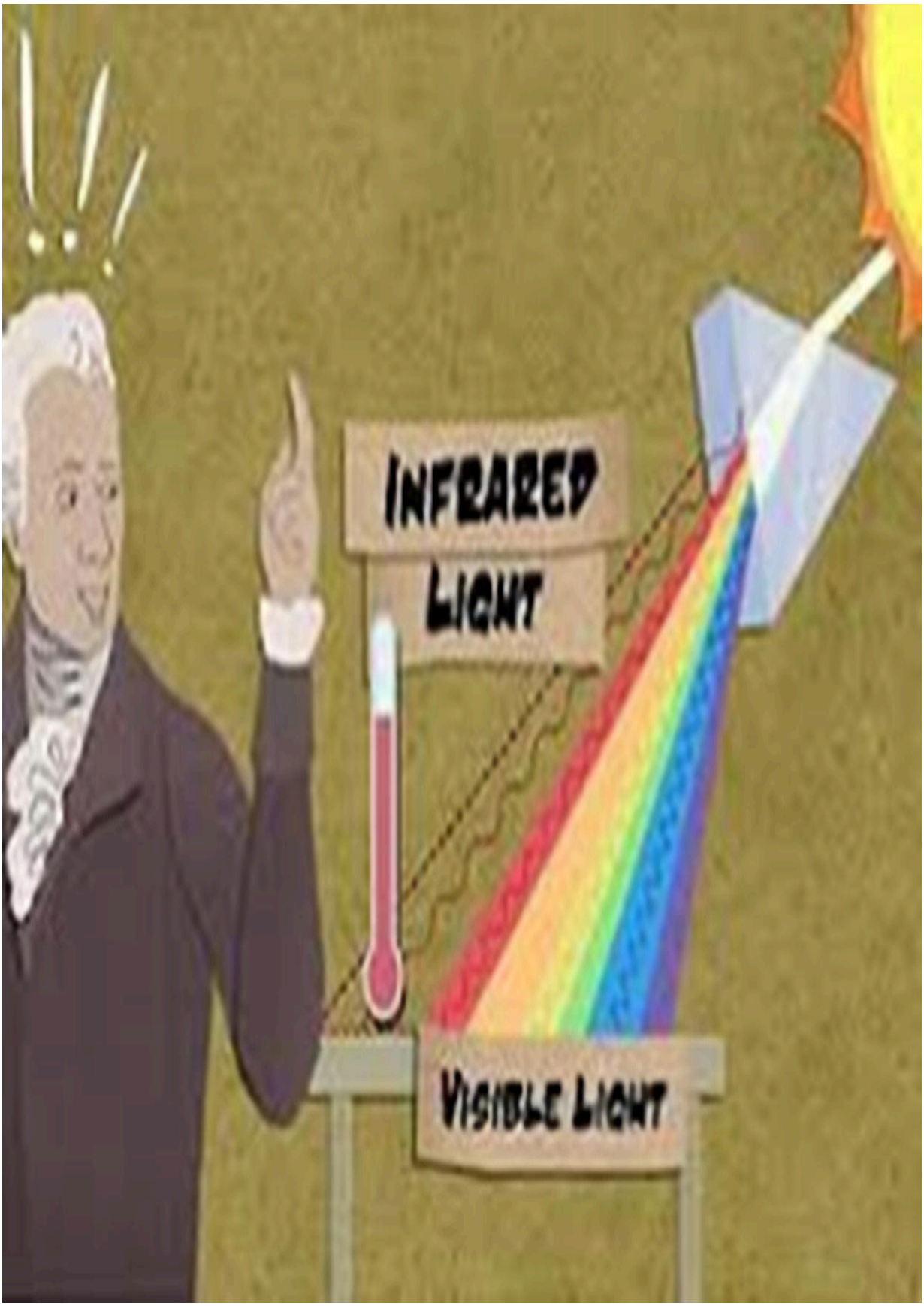

INFRARED
LIGHT
VISIBLE LIGHT

Figure 4.1 Hershel's experiment consisted of placing a thermometer beyond the red light and measuring the heat of the "invisible" radiation.

We have already seen in Chapter 1, Eq. (1.1), that frequency, *f*, and wavelength, λ, are related by the velocity of the wave, or

$$f = \frac{v}{\lambda} \tag{4.1}$$

The frequency, *f*, is measured in Hertz and a Hertz (named after Heinrich Rudolf Hertz [1857–1894; Figure 4.3], a German physicist for his work on electromagnetic waves) is the number of waves per second.

Einstein related the energy of a photon to its wavelength, Eq. (1.5), which I now rewrite using electron-volts (eV):

$$E = \frac{hc}{\lambda e} \tag{4.2}$$

where *E* is the energy of the light in electron-volts (eV), *h* is Plank's constant (6.63×10^{-34} m^2 kg s^{-1}), *c* is the speed of light (3×10^8 m s^{-1}), *e* is the electronic charge (1.6×10^{-19} coulombs), and λ is the wavelength (m). Notice that all the parameters are constants except for λ.

By replacing the constants in Eq. (4.2) by the value of the three constants we get a very simple relation between the energy of the electromagnetic waves in electron volts, *E*, and the wavelength, λ:

$$E = \frac{6.63\times10^{-34}\times3\times10^{8}}{10^{-6}\times1.6\times10^{-19}\times\lambda} = \frac{1.99\times10^{-25}}{1.6\times10^{-25}\times\lambda} = 1.24/\lambda \qquad (4.3)$$

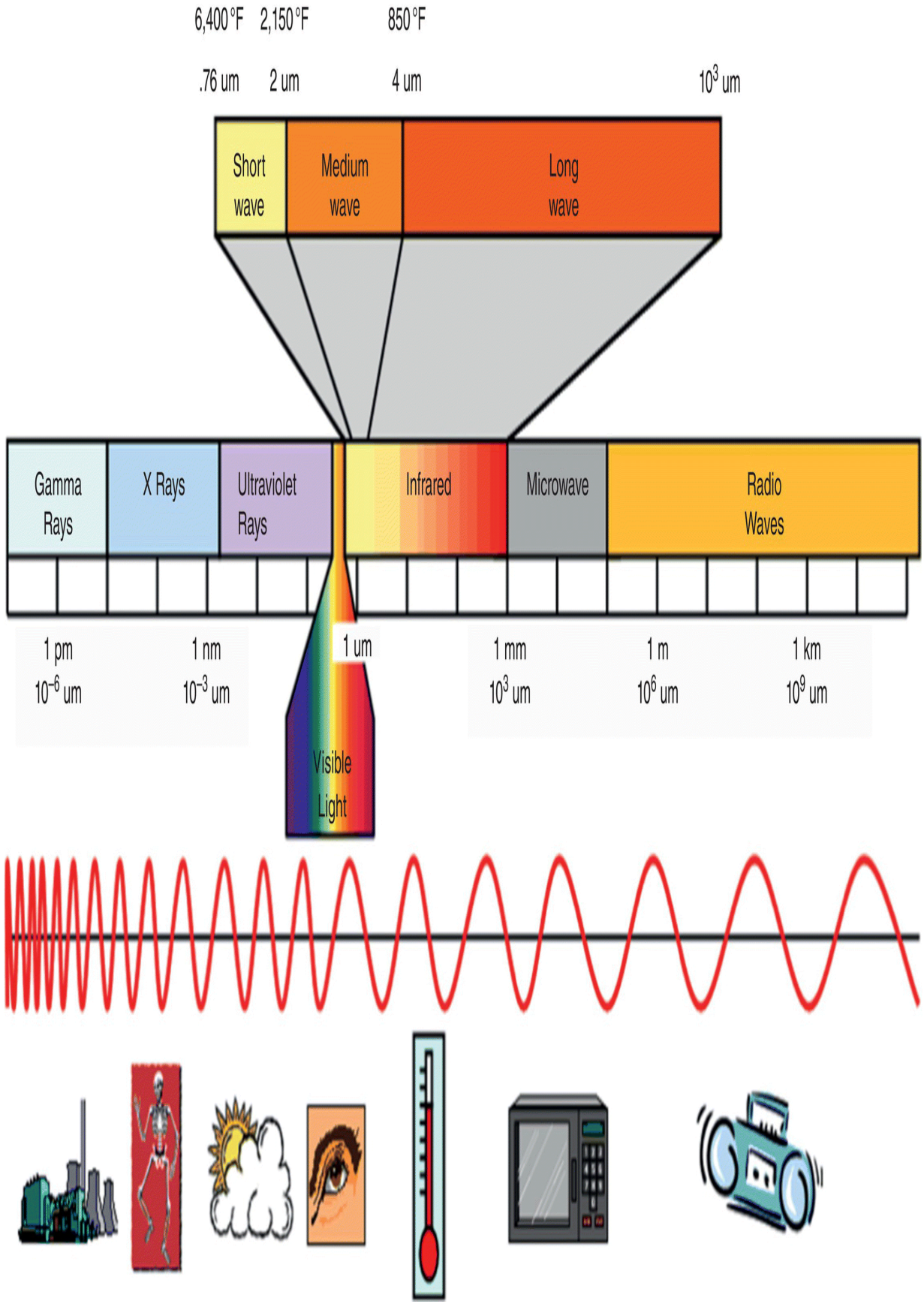

6,400 °F
2,150 °F
850 °F
.76 um
2 um
4 um
10^3 um
Short wave
Medium wave
Long wave
Gamma Rays
X Rays
Ultraviolet Rays
Infrared
Microwave
Radio Waves
1 pm
10^{-6} um
1 nm
10^{-3} um
1 um
1 mm
10^3 um
1 m
10^6 um
1 km
10^9 um
Visible Light

Figure 4.2 The entire radiation spectrum goes from gamma to radio waves, and the visible and the infrared ranges are a tiny portion of the entire radiation spectrum.

Figure 4.3 Heinrich Rudolf Hertz, who studied electromagnetic waves, was rewarded by having the units of frequency named after him.

Source: https://en.wikipedia.org/wiki/Heinrich_Hertz#/media/File:Heinrich_Rudolf_Hertz.jpg.

I'd like to take a look at the units of these variables to see if the relationship makes sense. So, let me do just that for you.

$$E = \frac{\text{joules} - \text{seconds} \times \text{meters} / \text{seconds}}{\text{meters} \times \text{coulombs}} = \frac{\text{joules}}{\text{coulombs}} \tag{4.4}$$

You can see that the seconds in the numerator cancel out and the meters in the numerator cancel the meters in the denominator and I am left with Joules (unit of energy) per coulomb (to convert energy to electron-volts), which is what I want. But be careful, the λ in Eq. (4.3) is in μm.

4.2 What Our Eyes Can See

It is very interesting to look also at the transmission of the atmosphere and the sun's radiation. The sun has a surface temperature of 5788 K, or about 10 000 °F, and at that temperature it generates radiation at different frequencies given by the black body radiation formula (see Appendix 4.2). The solid red line in Figure 4.4 shows the sun's radiation as a function of wavelength. The *x* axis of Figure 4.4 goes from zero to 2.5 μm. Notice that the sun emits the highest radiation precisely at the range of wavelengths we can see.

Figure 4.5 shows the opposite of the transmission of the atmosphere, that is, the opacity of our atmosphere. The *x* axis of this plot is now logarithmic. Notice how opaque our atmosphere is, except in the visible, infrared, and portions of the microwave and radio regions. In the infrared region, from 1 to 15 μm, the transmission is not as good but still a lot of radiation goes through. Isn't this interesting? Evolution has created an organ, our eyes, precisely tuned to where the sun has the highest radiation and the earth's atmosphere is almost transparent, which is a nice coincidence.

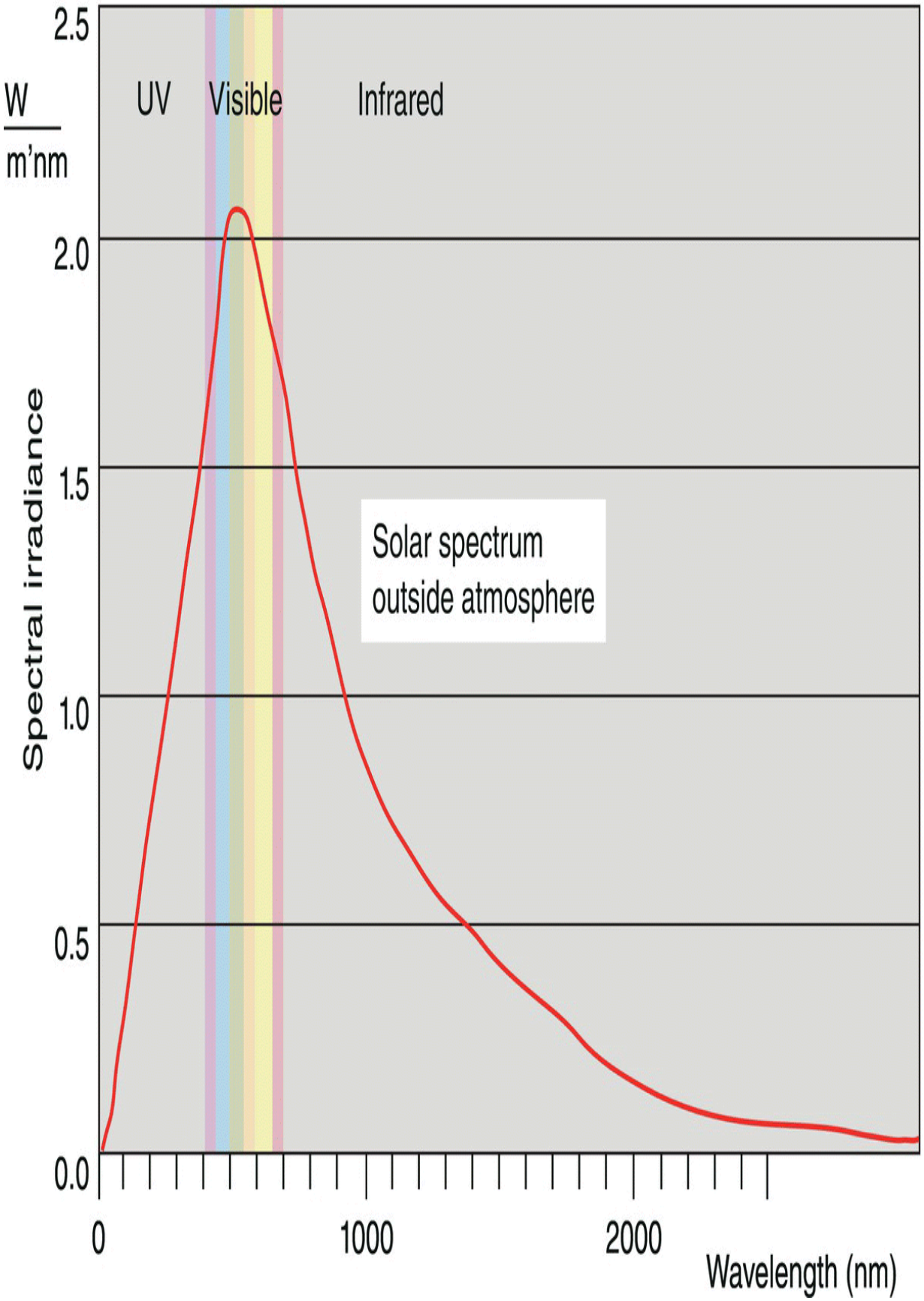

2.5
W
m'nm
UV
Visible
Infrared
2.0
1.5
1.0
0.5
0.0
Spectral irradiance
Solar spectrum
outside atmosphere
0
1000
2000
Wavelength (nm)

Figure 4.4 The sun's radiation spectrum is strongest in the range of wavelength, that is, colors, that our eyes can see.

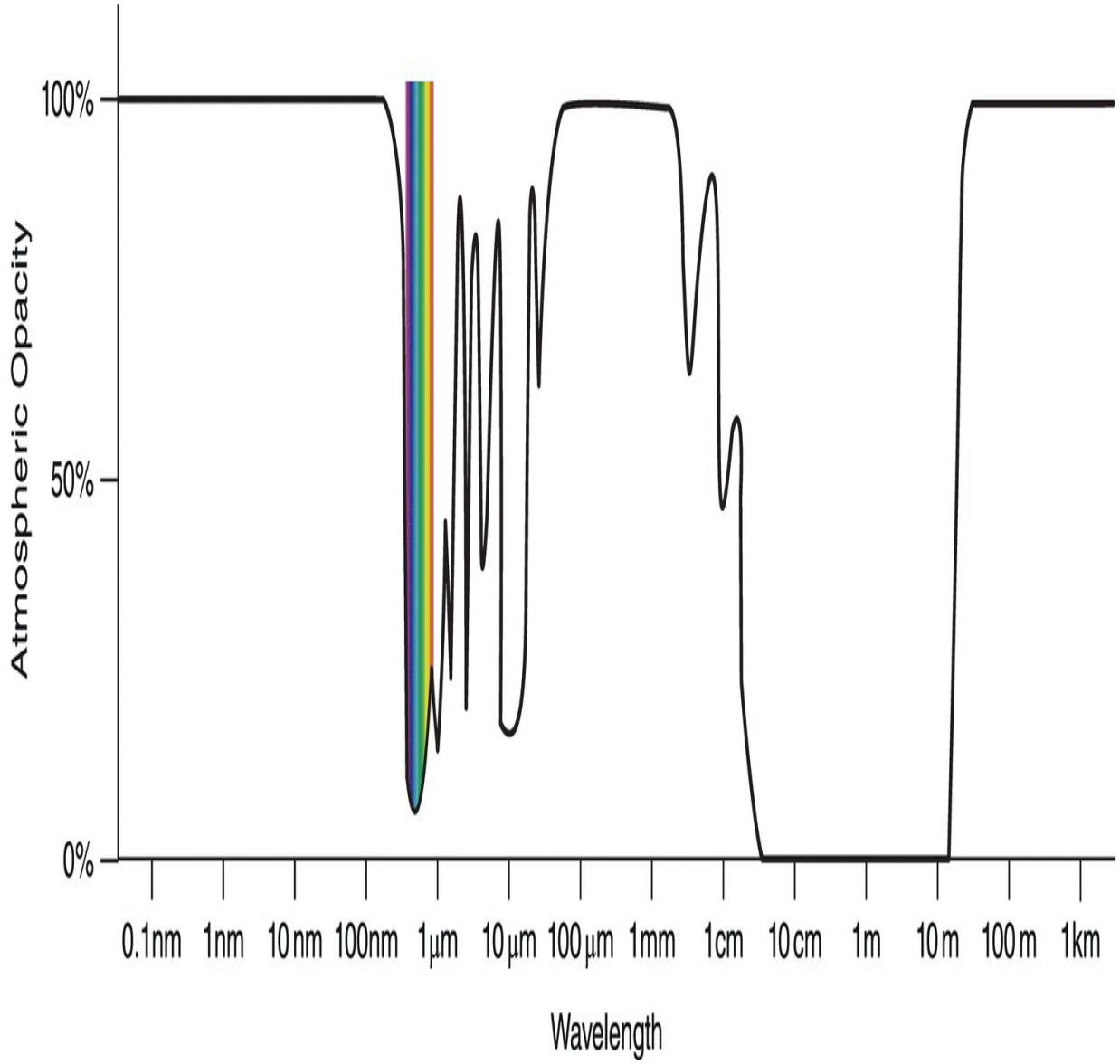


Figure 4.5 The earth's atmosphere is opaque except in the visible range and in portions of the infrared and radio ranges.

4.3 Infrared Applications

Infrared detectors have lots of applications, mostly based on the ability to "see" temperature differences. Figure 4.6 shows on the left a photograph of a man holding a scorpion. On the right is the same photograph taken with an infrared camera. Using software, we assign "colors" to different

temperatures (see the scale on the right of the infrared photograph) so we can “see” the different temperatures. By looking at the photograph on the right we observe that the scorpion is a cold body creature. Take a look at the details of the person holding the scorpion. The parts of the t-shirt that touch the body are warmer than the folds separated from the body.

Figure 4.6 Visible and infrared photographs comparing the cold body of a scorpion to the warm body of the man holding it. The scale at the right shows the corresponding temperature of each color.

Figure 4.7 On the right the man hides his arm with a plastic bag. The arm is fully visible in the infrared photograph on the left.

Source: https://upload.wikimedia.org/wikipedia/commons/9/9b/Human-Visible.jpg (left); https://en.wikipedia.org/wiki/Thermography#/media/File:Human-Infrared.jpg (right).

This is just an example. Infrared can make visible the level of water or oil in a tank by "seeing" the change in temperature in the metallic walls. People or animals can hide their images behind bushes, but they cannot hide the heat of their bodies. They are clearly visible in the infrared (Figure 4.7). Engineers can look at a house with an infrared camera and determine where there are heat leaks and therefore where insulation needs to be installed or improved (Figure 4.8). Art custodians can reveal a hidden painting under a visible one and help determine if the painting is or is not an original. All of you have several infrared clickers on the coffee table in front of the TV that control the channels and other electronic gadgets.

One of the scientific uses of infrared detectors is in astronomy. (Disclosure: I lead the engineering team that developed some of the infrared detectors for the infrared instruments in the Spitzer and the future Jack Webb space telescopes.) Looking at the sky using different wavelengths, the astronomer learns a lot about the universe. Take a look at Figure 4.9. On the left I show one of the first photographs of the Hubble telescope, the Eagle nebula. When we take a photograph of the same nebula in the infrared, all of a sudden myriad stars show up, stars that were obscured in the visible range by the dust and gas around the Eagle nebula.

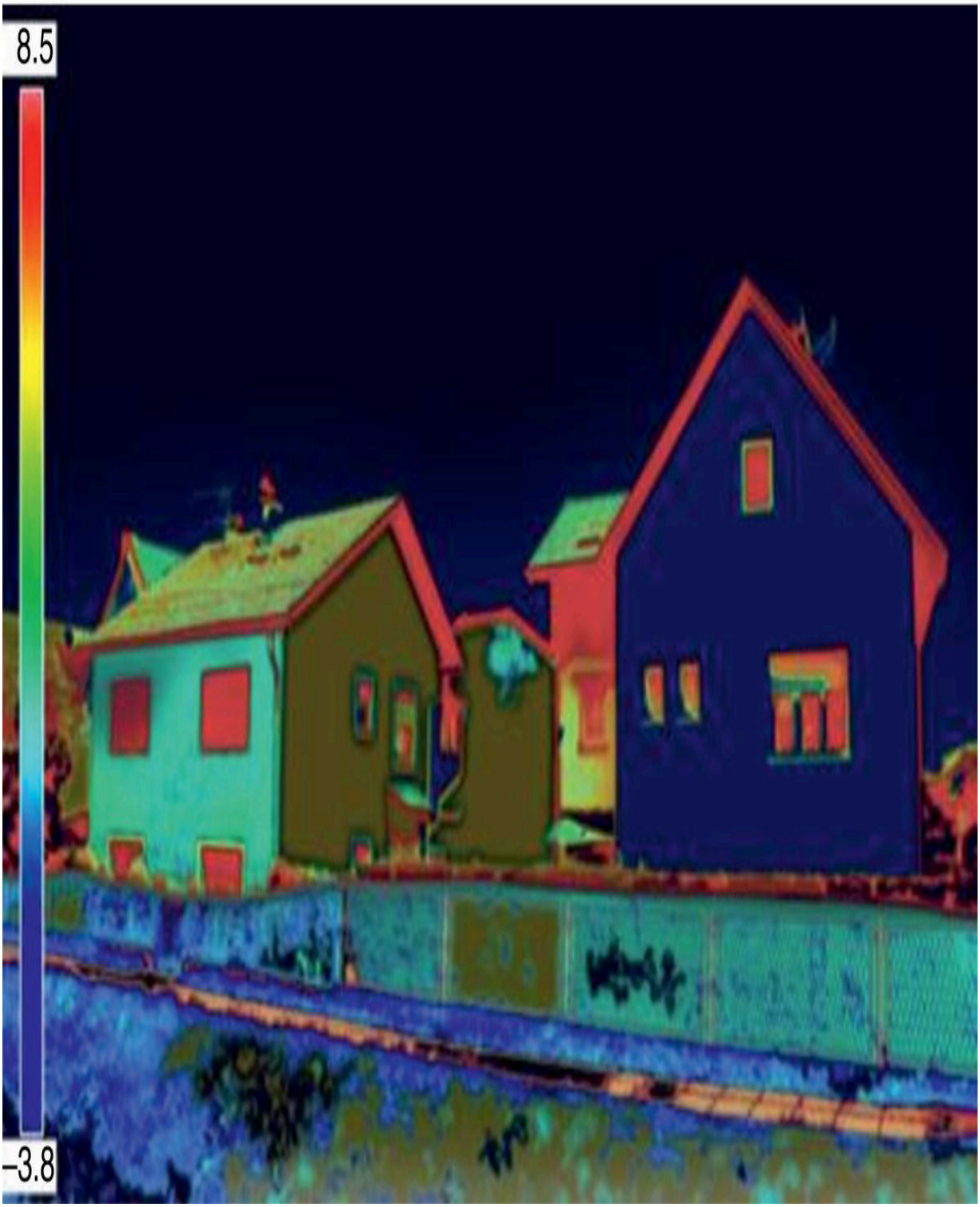


Figure 4.8 This infrared image of houses shows where heat is lost due to lack of proper insulation.

Source: https://www.123rf.com/stock-photo/39603239.html?oriSearch=image+id+39603239&sti=lvim4r0ejxisbgpt3s|.

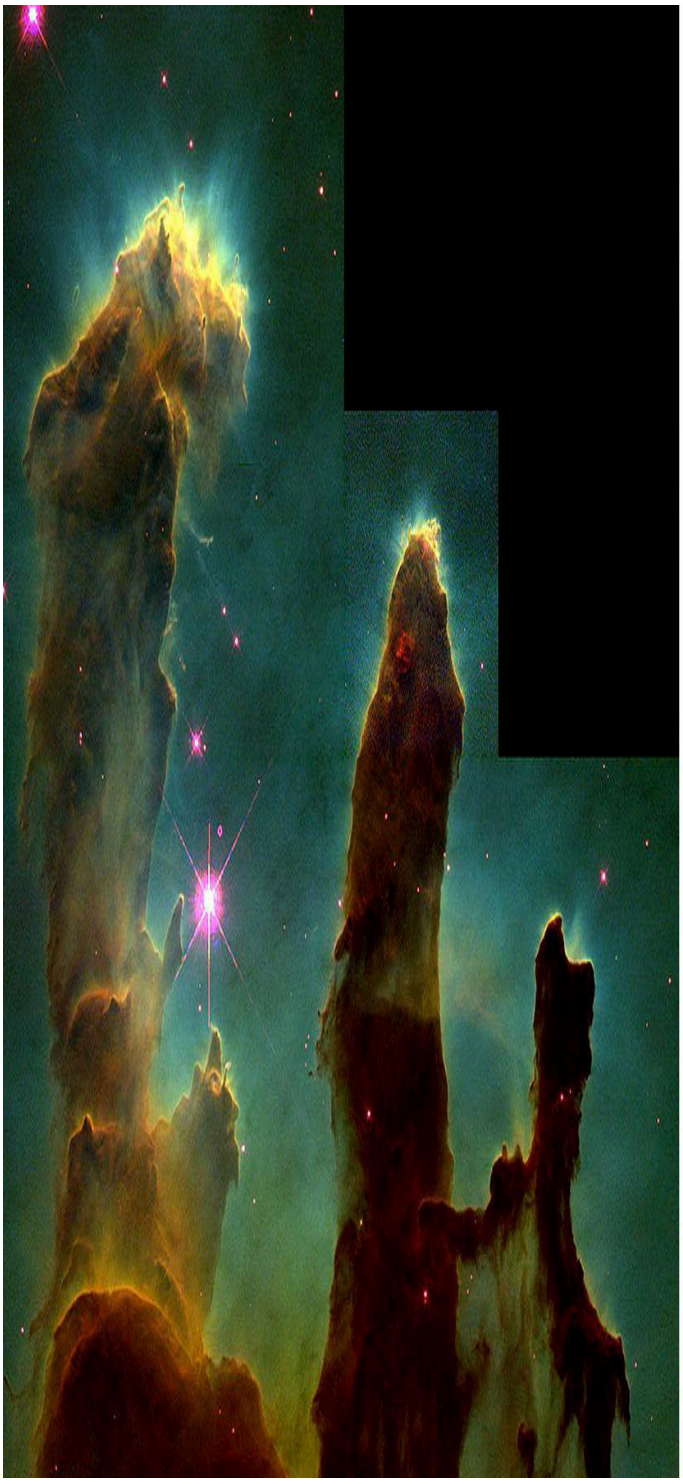


Figure 4.9 The Eagle nebula captured by the Hubble telescope using the visible range (on the left) and the same photograph taken in the infrared (on the right). The infrared image shows details that are obscured in the visible range by interstellar dust.

Source: http://www.spitzer.caltech.edu/images/1517-ssc2005-23b1-Hubble-Image-of-the-Eagle-Nebula (left); https://www.google.com/search?q=nasa+images+eagle+nebula&client=safari&rls=en&sxsrf (right).

4.4 Types of Infrared Radiation

After this digression on the use of infrared devices, let's go back to the physics of semiconductors and see how they make infrared detection possible.

As the wavelength gets longer and longer, or the frequency shorter and shorter, the energy of the infrared photons gets smaller and smaller. Look back at Eqs. (4.2) and (4.3). The wavelength of what we call near-infrared radiation (NIR) ranges from 0.75 to 2.5 μm, the mid-infrared radiation (MIR) from 2.5 to 6 μm, and the far-infrared radiation (FIR) from 6 to 15 μm. Above those ranges comes the extreme infrared radiation (XFIR) that goes from 15 all the way up to 1000 μm. The semiconductor infrared detectors are not useful at these very long wavelengths and you will see why.

Using the relationships at the beginning of this chapter, $E = 1.24/\lambda$ (Eq. 4.3) and $f = c/\lambda$ (Eq. 4.1), we can list the frequencies, wavelengths, and energies of the infrared spectrum, see Table 4.1. The energy of the radiation keeps getting smaller as the frequency decreases.

Table 4.1 Frequency, wavelength, and energy of photons in the four infrared ranges.

Radiation band	Frequency (1 s^{-1})		Wavelength (μm)		Energy (eV)	
	Beginning	Ending	Beginning	Ending	Beginning	Ending
NIR	4.0×10^{15}	1.2×10^{15}	0.75	2.5	1.65	0.5
MIR	1.2×10^{15}	0.5×10^{15}	2.5	6.0	0.5	0.2
FIR	0.5×10^{15}	0.2×10^{15}	6.0	15.0	0.2	0.08
XFIR	0.2×10^{15}	0.003×10^{15}	15.0	1000	0.08	0.001
Microwave	0.003×10^{15}		1000		0.001	

4.5 Extrinsic Silicon Infrared Detectors

The energy of infrared radiation is too low to kick an electron from the valence band to the conduction band of intrinsic, pure, no-impurities, silicon (remember we needed 1.12 eV to free an electron). For many applications, especially in the astronomic field, we want to see in the MIR. The only way to do this is to use doped semiconductors with impurities close to the conduction band so that photons in the infrared range have enough energy to kick some of these electrons into the conduction band. We do not want any free electrons due to the thermal energy (remember practically all the electrons in the donor band move to the conduction band at room temperature, Sections 3.4 and 3.5). Therefore, we start by cooling the detectors to extremely low temperatures as close as possible to the temperature of liquid helium, 4 K (−270 °C or −452 °F). This low temperature ensures that all the electrons occupy the lowest possible energy, as I show in Figure 4.10.

We like to use arsenic, As, to detect photons in the MIR region of the spectrum. We saw in the previous chapter that the energy to free, that is, ionize, an electron from an As atom is 0.054 eV, a much lower energy than the 1.12 eV needed to do the same in the intrinsic silicon. We use arsenic as the preferred doping gas because it is easier to introduce in very small amounts while we fabricate the silicon. Figure 4.10 shows the energy gap with all the fifth electrons of the As atoms occupying the donor levels close to the conduction band when the doped Si is very close to absolute zero (0 K).

When an infrared photon with an energy greater that 0.54 eV strikes the As doped semiconductor, the photon has sufficient energy to free one of the electrons that are sitting in the donor levels very close to the conduction band. So, if we detect a free electron in the conduction band, we know that, since the electrons are frozen at the low temperature, a photon with an energy higher than 0.054 eV has hit the crystal. The more electrons I can find in the conduction band the more photons I know have been absorbed by the detector. Figure 4.11 shows the energy transfer of the photon with an energy larger than 0.054 eV to an electron.

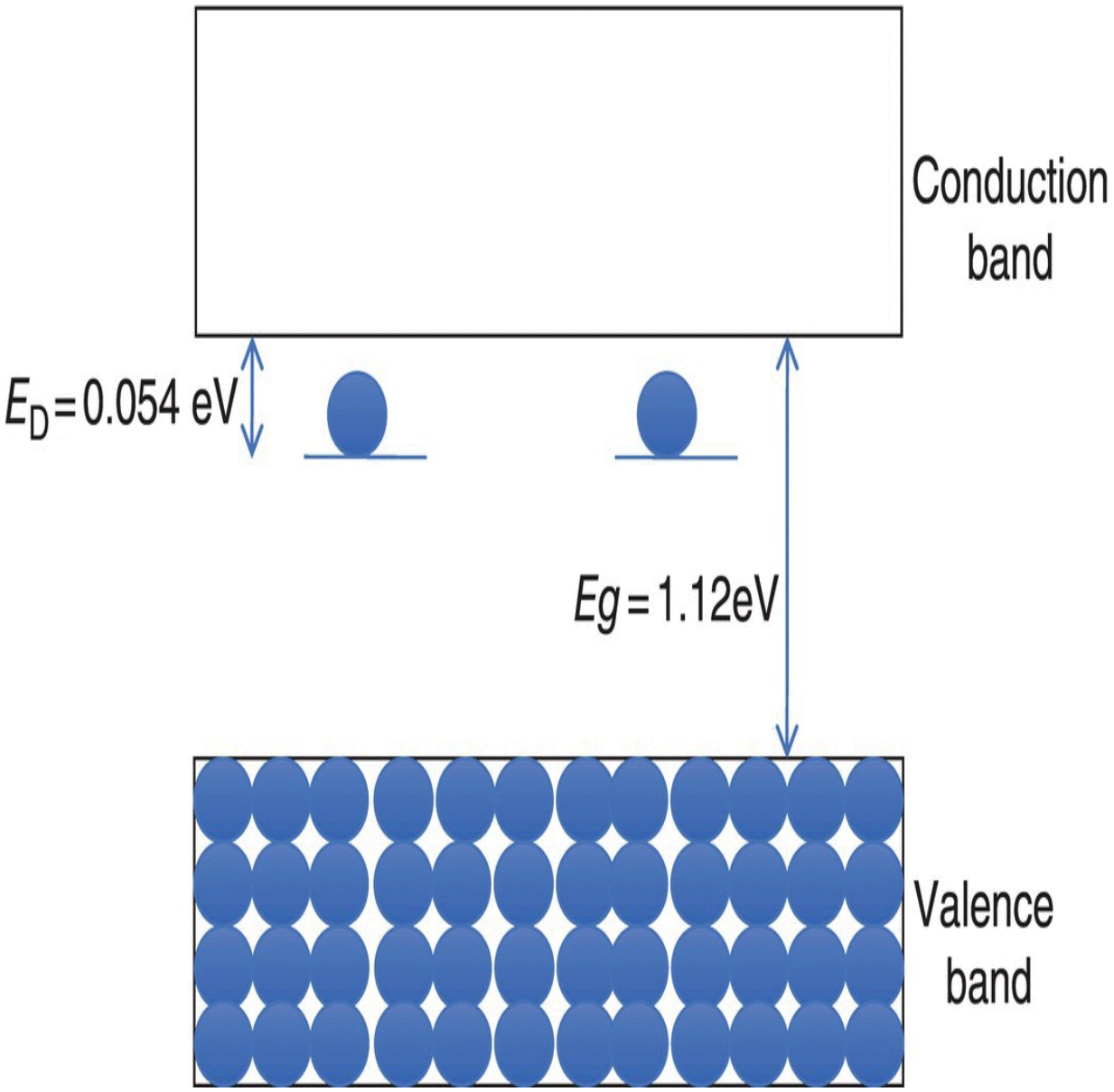


Figure 4.10 At very close to absolute zero all the electrons from the donor atoms occupy the lowest possible allowed energies, that is, the donor levels.

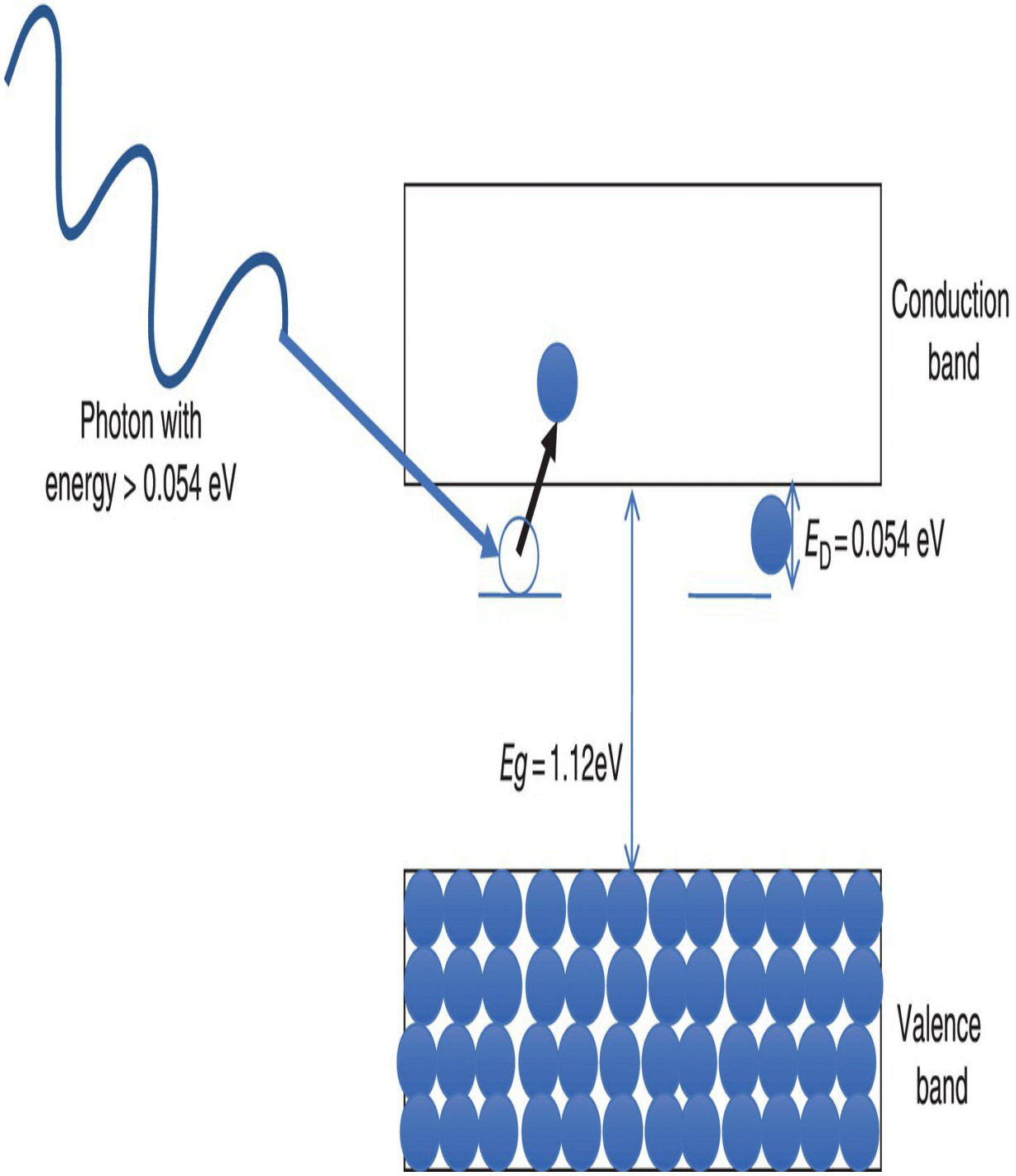


Figure 4.11 A photon with energy greater than 0.054 eV hits the As-doped silicon, sending an electron to the conduction band.

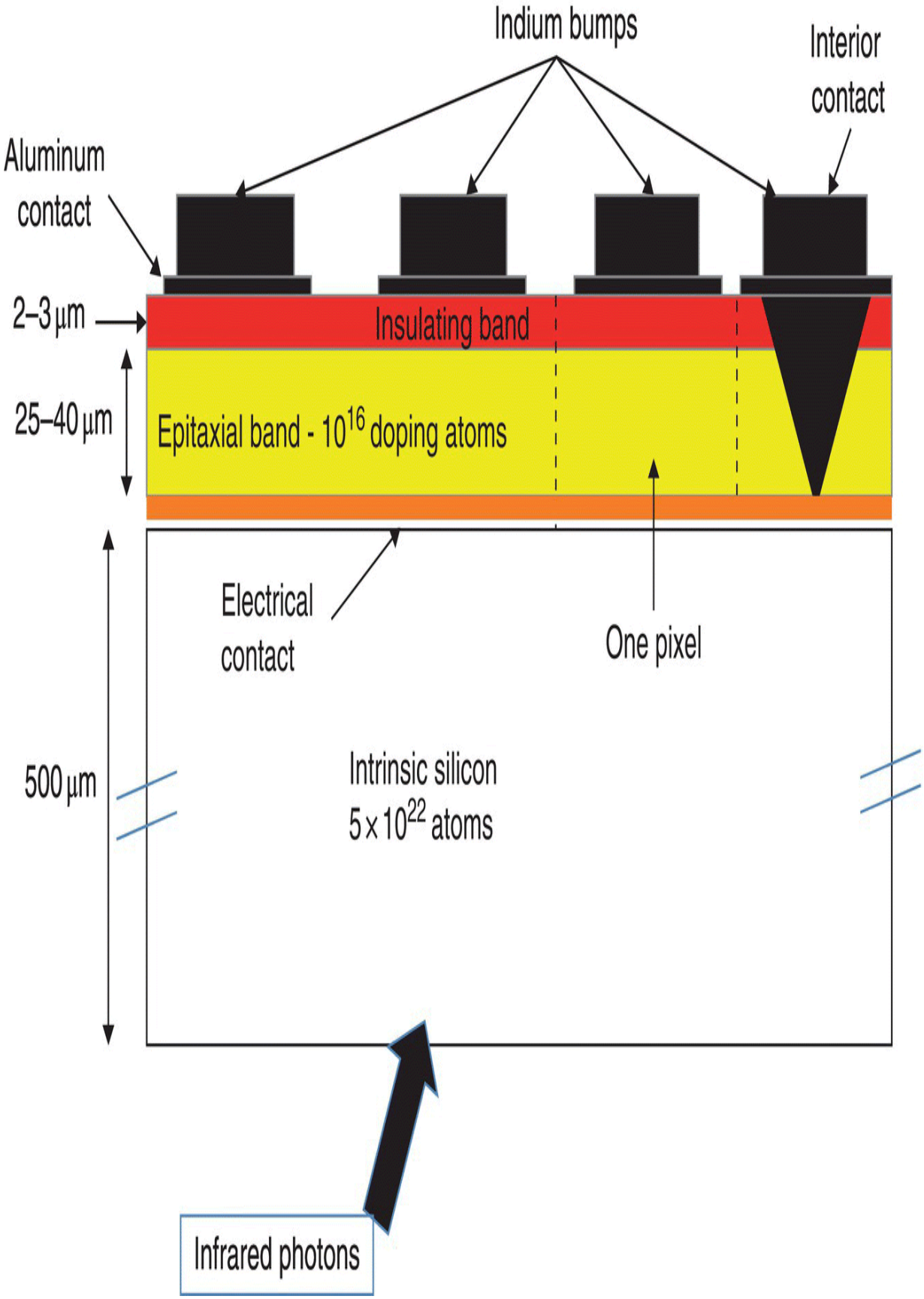

Indium bumps
Interior contact
Aluminum contact
2–3 µm
Insulating band
25–40 µm
Epitaxial band - 10^{16} doping atoms
Electrical contact
One pixel
500 µm
Intrinsic silicon
5×10^{22} atoms
Infrared photons

Figure 4.12 Cross-section of an arsenic doped infrared detector showing the substrate (white), internal contact (dark yellow), absorbing epitaxial layer (yellow), insulating oxide (red), and various contacts (black).

If you look at Table 4.1, you will realize that the arsenic impurities are able to detect any photon in the NIR and MIR ranges. You may ask what happens if visible light hits this doped silicon. Will the visible light mess up everything by depleting all the electrons from the donor band, overwhelming the conduction band? The answer is yes, so these infrared systems have filters in the optical path that do not let photons above a certain energy go through and thus prevent them from interfering with the photons we are interested in looking at.

Let me now explain the challenges involved in fabricating these devices. We call them extrinsic detectors because they use the doping levels instead of the gaps between conduction and valence bands. Figure 4.12 shows a cross-section of one of those detectors. Here is the list of the steps involved in fabricating these devices. (In Chapter 10 I go over many of the methods used to fabricate silicon devices.) Here I just mention the structure we need to fabricate a working extrinsic infrared detector.

The 500-μm thick intrinsic silicon (in white) is needed to support the very thin epitaxial layer that we grow on top of it. The photons hit the detectors from the bottom (there are lots of electronic elements on top of the detector that I do not show). We want the silicon substrate to be transparent so photons can go through without being absorbed. We want, therefore, the silicon substrate to be very pure, with no impurities or imperfections. We want the impurities in the substrate to be under 10^{13} atoms per cm^3. This is not easy. It means that we can only tolerate one impurity atom for every five billion silicon atoms.

On top of this intrinsic semiconductor we grow a very thin layer of highly doped silicon (the continuous dark yellow line). This is the internal conductive layer and serves as the detector's back contact. It has to be very thin so that not many photons are captured, and thus lost, as they cross the contact layer.

Now comes the critical area. We epitaxially (explained in Section 10.3) grow one silicon layer at a time over the silicon substrate (yellow), and as

we do that we add a very controlled amount of arsenic atoms using a gas like arsine (AsH_3) such that just about 10^{16} atoms of gas per cm^3 replace the same number of the silicon atoms (remember Figure 3.9). We do not want the As atoms to interact with each other. That epitaxial layer is between 25 and 40 μm thick. The thicker the layer the more photons we collect, but it is much more difficult to fabricate a thicker silicon epitaxial layer without any defects.

On top of the epiaxial layer we grow a very thin (1–2 μm) insulating band of SiO (in red).

We need to make a hole in the layers (the inverted triangle on the far right of Figure 4.12) so that we have an electrical contact to the conductive layer between the substrate and the epitaxial layer.

Finally, we fabricate the contacts (black). These are small ($30 \times 30\ \mu m^2$ or smaller) aluminum pads and on top we deposit an indium bump. The aluminum pads define the size of the pixels. We fabricate 1024×1024 (or 2056×2056) such pixels and the indium bump is used to make contact with another chip, the readout (I discuss this other electronic readout chip after I explain how the transistors and integrated circuits work, see Appendix 13.1). This readout chip, also with 1024×1024 inputs, contains all the electronics needed to capture the excited electrons from each detector pixel and process the information. Figure 4.13 shows a photograph of the contacts and indium bumps.

The final detector assembly is shown in Figure 4.14. The completed assembly consists of the detector material and the readout chip. The readout chip has the same number of input structures as detectors. Each detector is connected to its own input circuit by an indium bump. The infrared radiation hits the detector array, from the top in Figure 4.14, and each detector absorbs photons proportional to the intensity of the radiation at that location. The detector changes the photons into electrons that are read by the input structure of the readout array. The readout has a horizontal and a vertical multiplexer (Section 12.1) that select one cell at a time and send the information sequentially to a processor that interprets the data and creates an image.

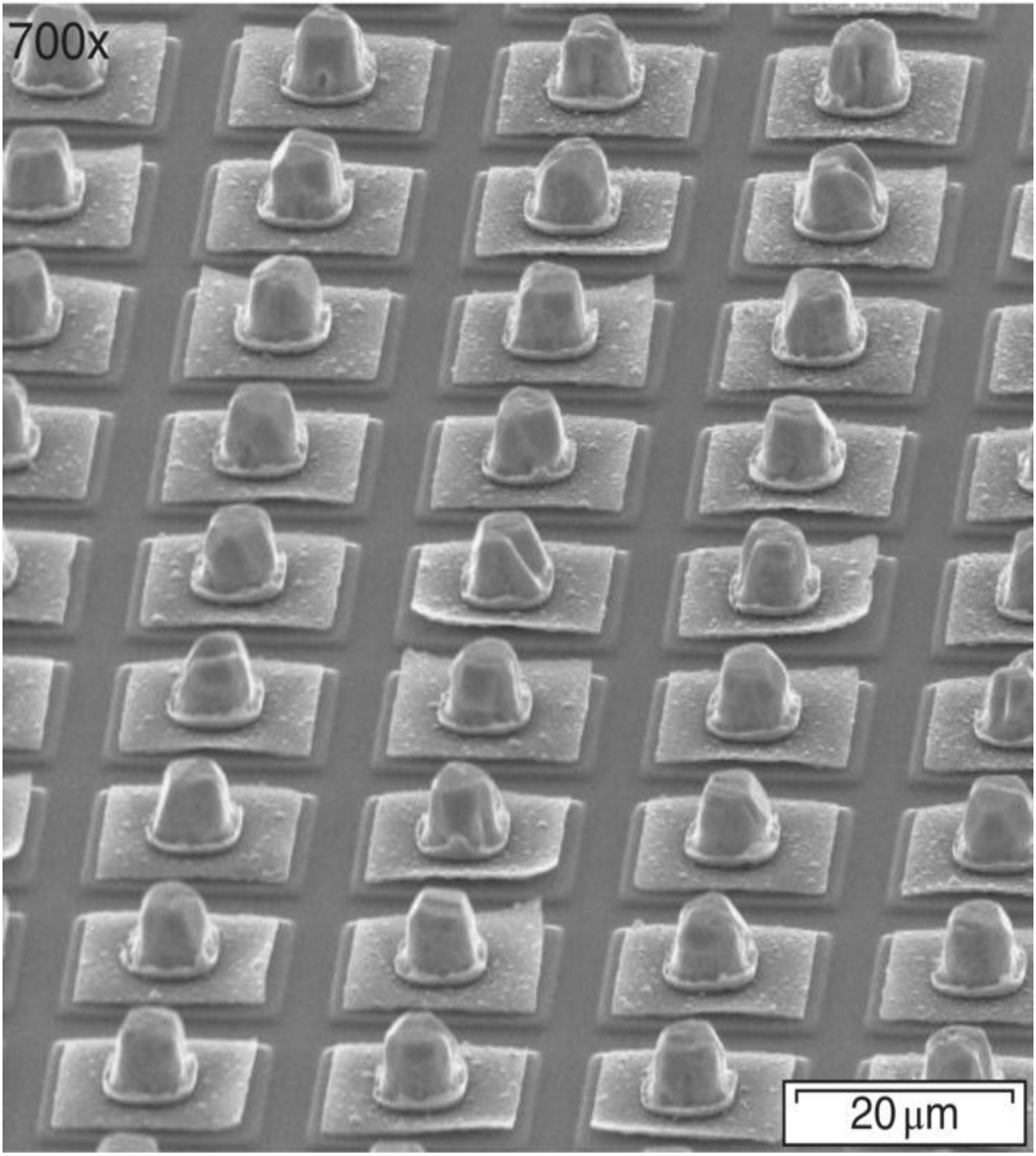


Figure 4.13 A photograph of the contacts and indium bumps that define and connect each one of the one million pixels to the appropriate input of an electronic chip.

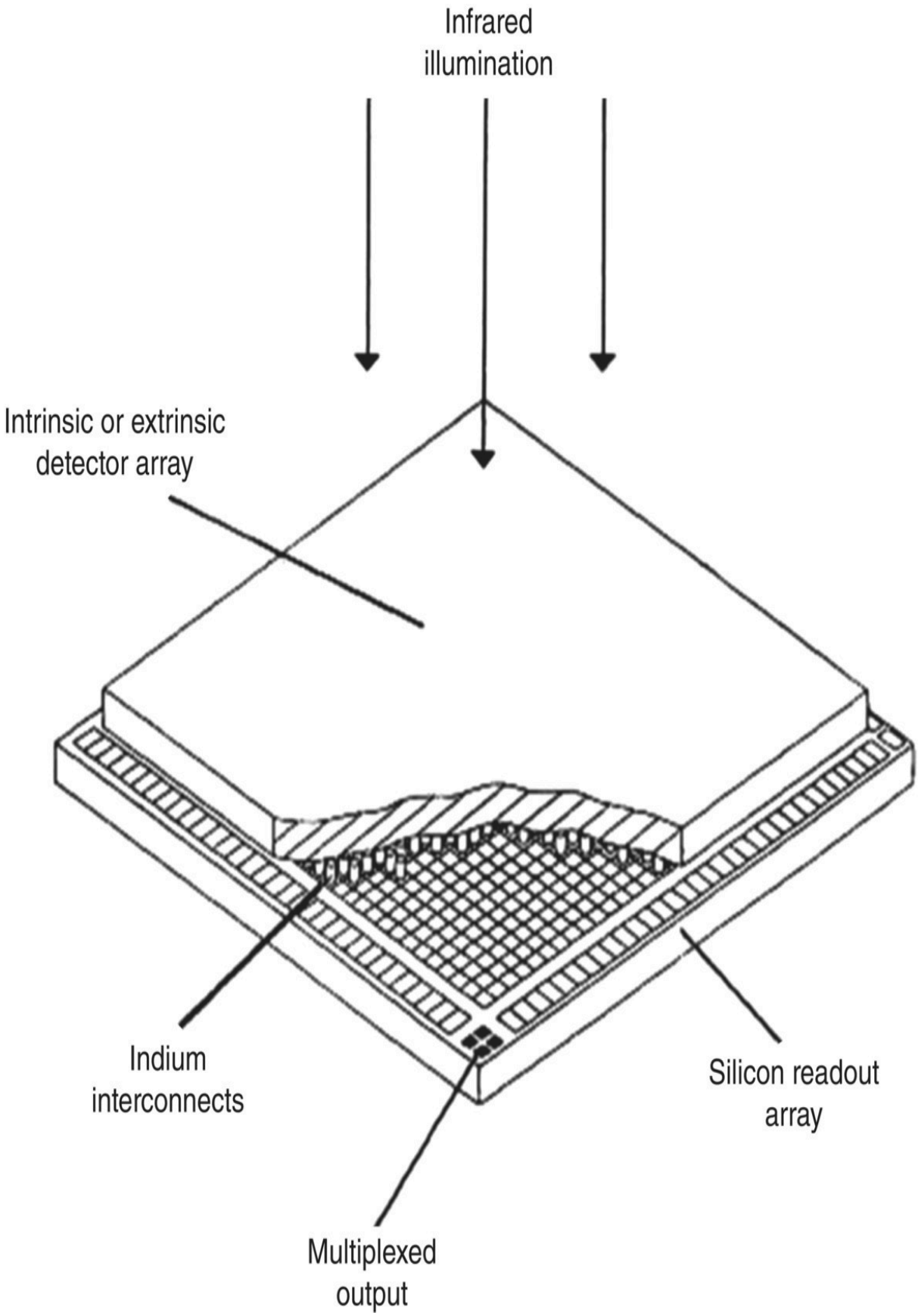

Infrared
illumination
Intrinsic or extrinsic
detector array
Indium
interconnects
Multiplexed
output
Silicon readout
array

Figure 4.14 A completed detector assembly with the detector array on top of the readout chip connecting each detector to one input cell via indium bumps.

The detectors that I have described here are the retinas of cameras or telescopes. The telescope is nothing more than an artificial eye. It has a supportive structure (cornea and ciliary body), a shutter (iris), optics (lens), environmental needs (aqueous and vitreous body), which in our case is the helium cooler, a radiation detector (retina), and indium bumps (optic nerve) that connects the output of the detector to the electronics and signal processor (brain).

These infrared devices and systems are now flying in the Spitzer astronomic observatory (I show one of the photographs from the Spitzer telescope in Figure 4.9) and in the near future, currently intended to be 2021, in the Jack Webb infrared astronomical observatory.

4.6 Intrinsic Infrared Detectors

The silicon doped with As is just a special type of semiconductor infrared detector. I started with it because is a nice example whose operation can be explained by the use of energy bands and the effects of impurities in semiconductors. More commonly used are the indium-antimonite (InSb) and the mercury-cadmium-tellurite (HgCdTe) detectors. Let us find out why.

HgCdTe can be grown in many combinations. As a matter of fact, in the infrared literature you will find that the formula is written as $Hg_{1-x}Cd_xTe$. This is because, depending on how much mercury and cadmium we have (notice that the sum of Ca + Hg equals the number of tellurium atoms), the energy gap changes. For example, if $x = 0.2$ (i.e. we have 20% cadmium and 80% mercury), the energy gap is $Eg = 0.09$ eV but if we increase this to $x = 0.6$, the band gap increases to 0.75 eV. This means that the energy gap is small enough that NIR and some of the MIR radiation has enough energy to free charges from the valence band to the conduction band without the need for impurities. These detectors act like an intrinsic semiconductor with a very narrow and adjustable energy gap.

The advantages of HgCdTe detectors are that they can work at higher temperatures, including at room temperature for NIR applications, and by changing the composition of mercury and cadmium we can tailor their operation to our desired wavelength. The main disadvantage is that they are much more difficult to fabricate with the level of purity and the few imperfections we need.

Another common material used for infrared detectors is indium antimonite, InSb. The energy gap of InSb is 0.18 eV at room temperature (300 K) and 0.23 eV at 77 K. Cooling a detector to 77 K is not difficult, as this is the temperature of liquid nitrogen. The advantage is that at 77 K the number of intrinsic electrons, n_i, is only 2.6×10^9 compared to 2×10^{16} at room temperature, which means that there are very few intrinsic charges and thus the photons dominate the creation of free charges. That is what we want. As you can expect, it is also much easier to fabricate InSb than HgCdTe. As a matter of fact, the new infrared Jack Webb astronomical telescope has huge panels of InSb infrared detectors. Figure 4.15 shows the hexagonal panels of the primary mirror, filled with InSb detector arrays. Each of the 18 hexagonal segments is 1.32 m (4.3 ft) in diameter, each segment contains about 1000 detector arrays, and each array is composed of 2048 × 2048 detectors.

Infrared looks at the "heat" of the objects. Astronomers look at the sky at night because the light of the day interferes with the faint light from the stars. Infrared detectors look at the heat of objects and at night the environmental heat is still there. That is why infrared astronomical observatories have 100–1000 times better sensitivity in space compared to those on earth and the lower temperature is also one way to turn off any heat sources.

4.7 Summary and Conclusions

In this chapter we have seen how the concepts of energy bands and doped semiconductors are sufficient to explain how infrared detectors work. In the process we have seen how intrinsic semiconductors using HgCdTe and InSb have lower energy gaps that allow low energy photons to ionize electrons from the valence bands and extrinsic detectors with impurity levels very close to the conduction bands that permit the

detection of infrared radiation with still lower frequencies and lower energies.

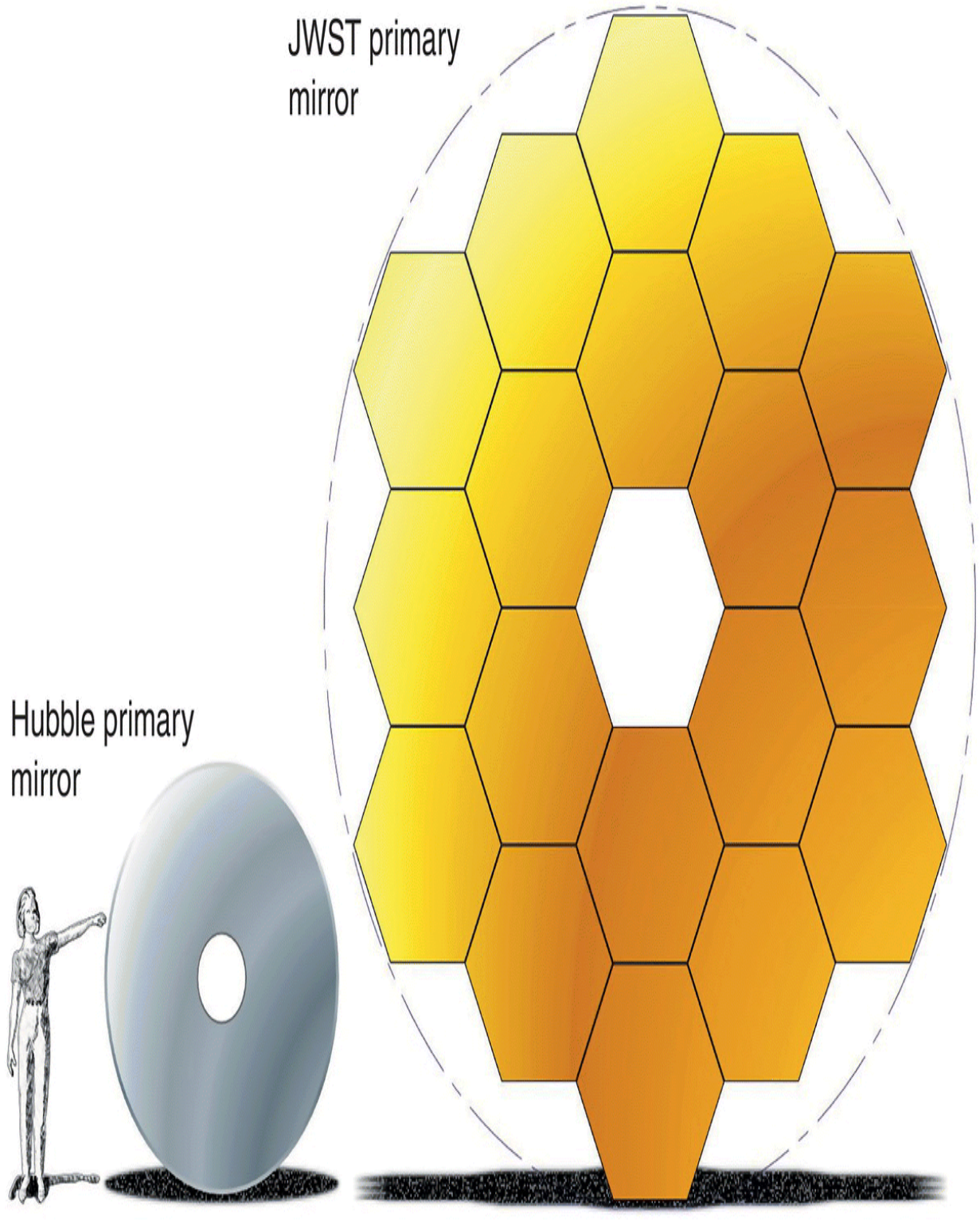

JWST primary mirror
Hubble primary mirror

Figure 4.15 The primary mirror of the Jack Webb telescope consists of very large hexagonal panels full of InSb infrared detector arrays. Compare the size of the mirrors to the height of a human and the Hubble primary mirror.

Source: https://www.jwst.nasa.gov/content/about/comparisonWebbVsHubble.html.

We will continue with semiconductor theory in next chapter and start learning about diodes, transistors, and other devices that we can build using semiconductors.

Appendix 4.1 Light Diffraction

You will have seen many times that if you put a stick into water, the stick seems to bend. The reason for this is that the velocity of light is different in air (almost vacuum) than in water (Figure 4.16). The right-hand side of Figure 4.16 shows what happens when a beam of light hits the surface of a transparent material like water. There is a reflected ray with a reflected angle identical to the angle of the incident wave, and a refracted ray with a different angle depending on the different optical properties and index of refraction of the two media (e.g. air and water).

One way of visualizing why a ray bends is to look at the right-hand side of Figure 4.16. A beam of light consists of a bundle of rays. When the first rays in the air hit the water, they slow down. The speed of light in water is 30% slower than the speed of light in a vacuum, or air. When the next array hits the water, the first ray has moved about 33% slower and so on with the other rays. This forces the beam to bend. Notice that the separation between the peaks (I show them as thinner lines perpendicular to the rays) is smaller in water than in air. The distance between peaks is by definition the wavelength. We can now say:

For reflection, the angles of the beams, θ_1 and θ_2, are equal, or

$$\theta_1 = \theta_2 \tag{4.5}$$

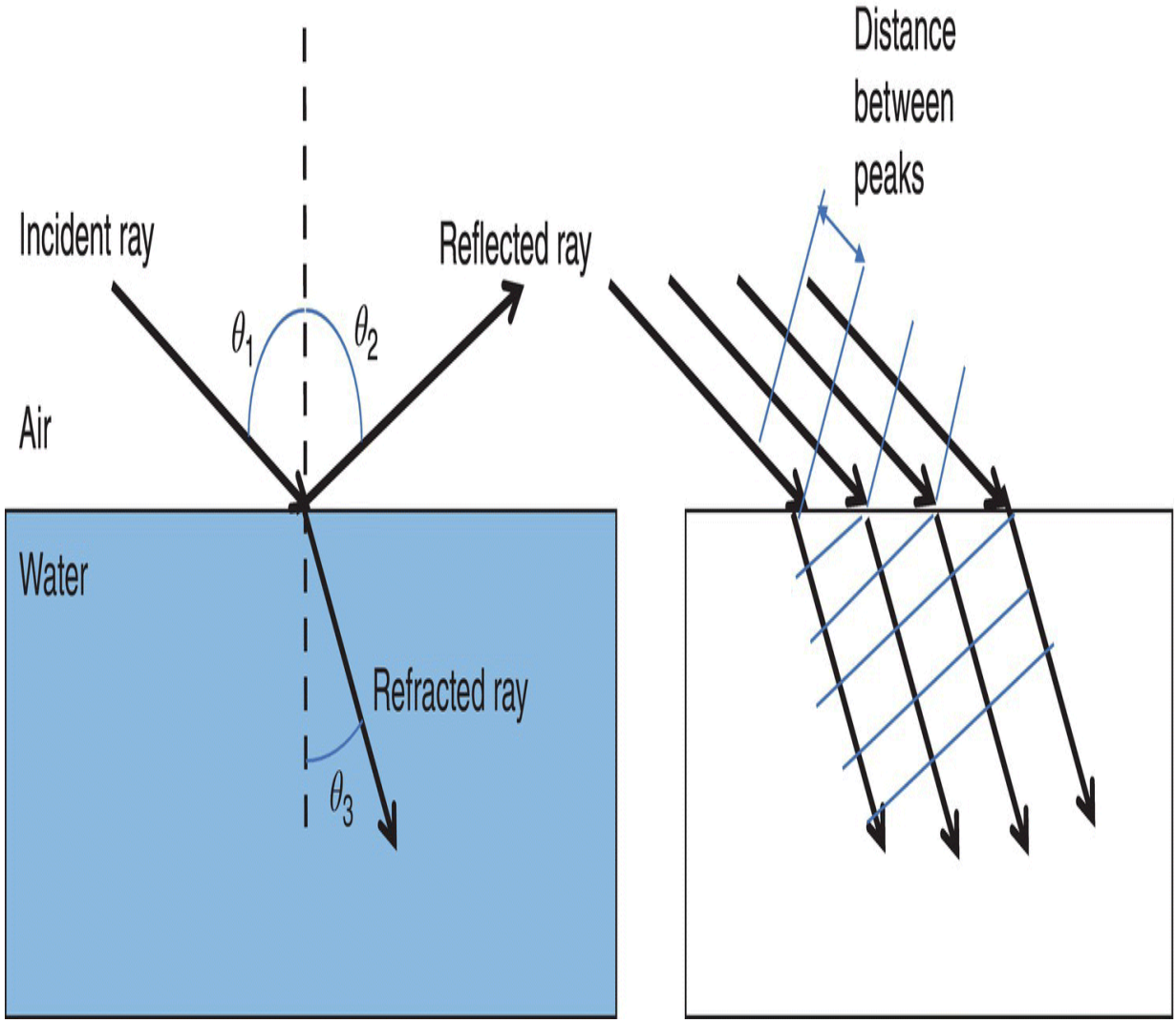


Figure 4.16 The reflection and refraction of light as it moves from air to water due to the different light velocities in the different media.

But for refraction the angles of the rays in the two media are given by the ratio of the sines of the respective angles, or

$$\frac{\sin\theta_1}{\sin\theta_3} = n_w \tag{4.6}$$

where n_w is the index of refraction of water, which happens to be 1.33. The index of refraction is different for different materials and it is not the same at all frequencies. The light interacts with the atoms in the materials and the atoms absorb and re-emit the light depending on the optical density of the material and these interactions are different for

different frequencies. The atoms of some materials can hold energy for a longer time than other materials. For crown glass, the index of refraction is 1.509 for red light and 1.521 for violet light. The result is that the violet light bends more than the red light as it crosses the prism and therefore the light colors are separated (Figure 4.17).

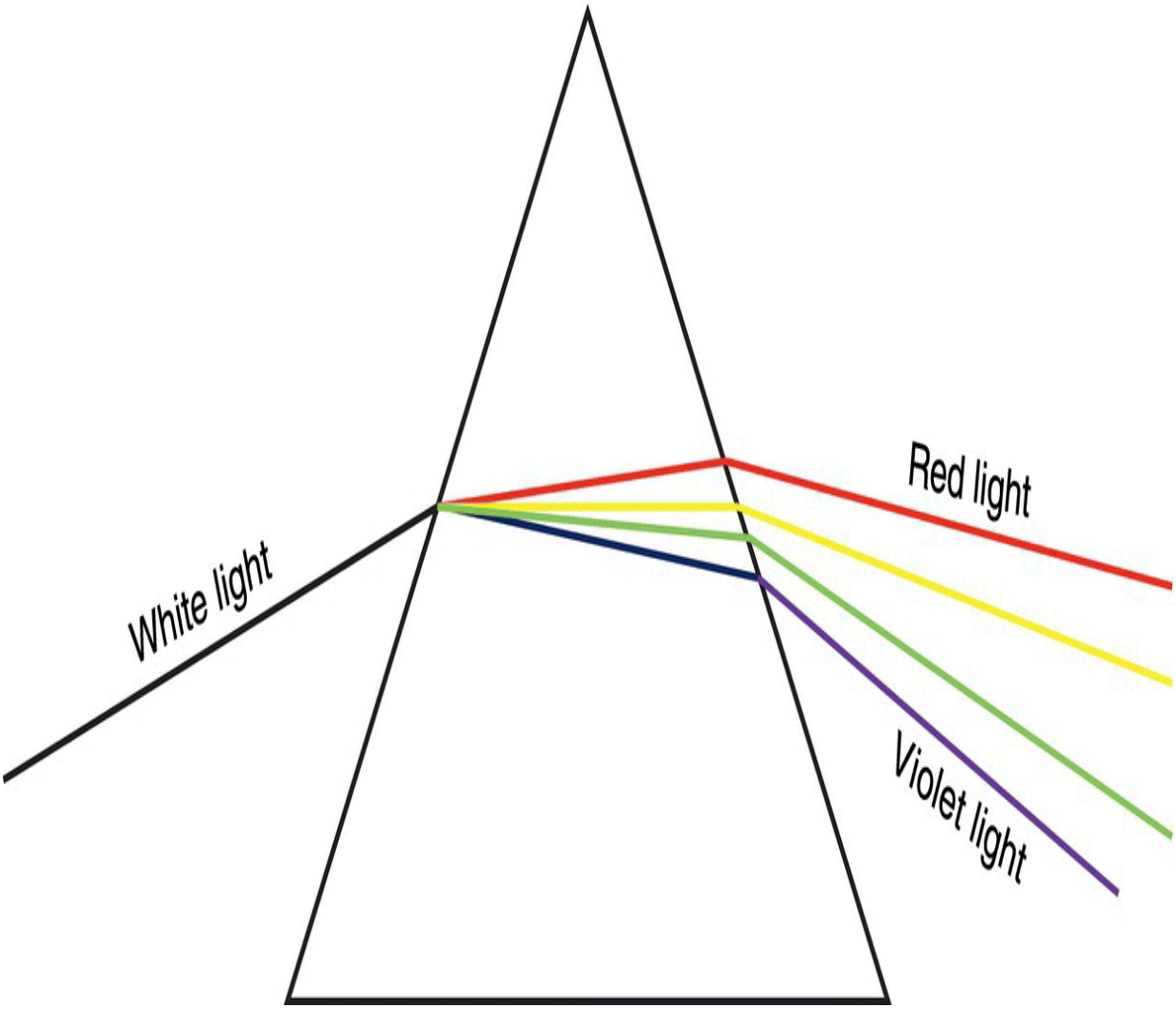


Figure 4.17 Light dispersion as it crosses a prism, separating the different colors.

Appendix 4.2 Blackbody Radiation

Scientists have studied radiation as a function of frequency and wavelength. In 1860, Gustav Kirchhoff (1824–1887; Figure 4.18) coined the term "blackbody" to describe a source that absorbs or emits all the

radiation. We see colors because part of the spectrum is being absorbed by objects. We can say that an orange absorbs all the colors except red. The red color is reflected from the peel of the orange and goes to our eyes. All the other colors are absorbed inside the orange peel. A blackbody is one that absorbs absolutely *all* radiation, of all frequencies, including, of course, colors. Nothing comes out; it is completely black, a humanly manufactured "black hole."

There have been many attempts to explain the amount of radiation that a blackbody could generate. The problem is that all the classical explanations using classical thermodynamics concluded that a blackbody would emit an infinite amount of radiation, which, of course, is not possible.

Max Planck (1858–1947; Figure 4.19) concluded that classical theories could not explain radiation as a function of frequency. Using quantum mechanics, he assumed that the energies were quantized, as we saw also in the Bohr atom. He came up with the equation

$$W = \frac{2\pi hc^2}{\lambda^5} \times \frac{1}{e^{-ch/kT} - 1} \tag{4.7}$$

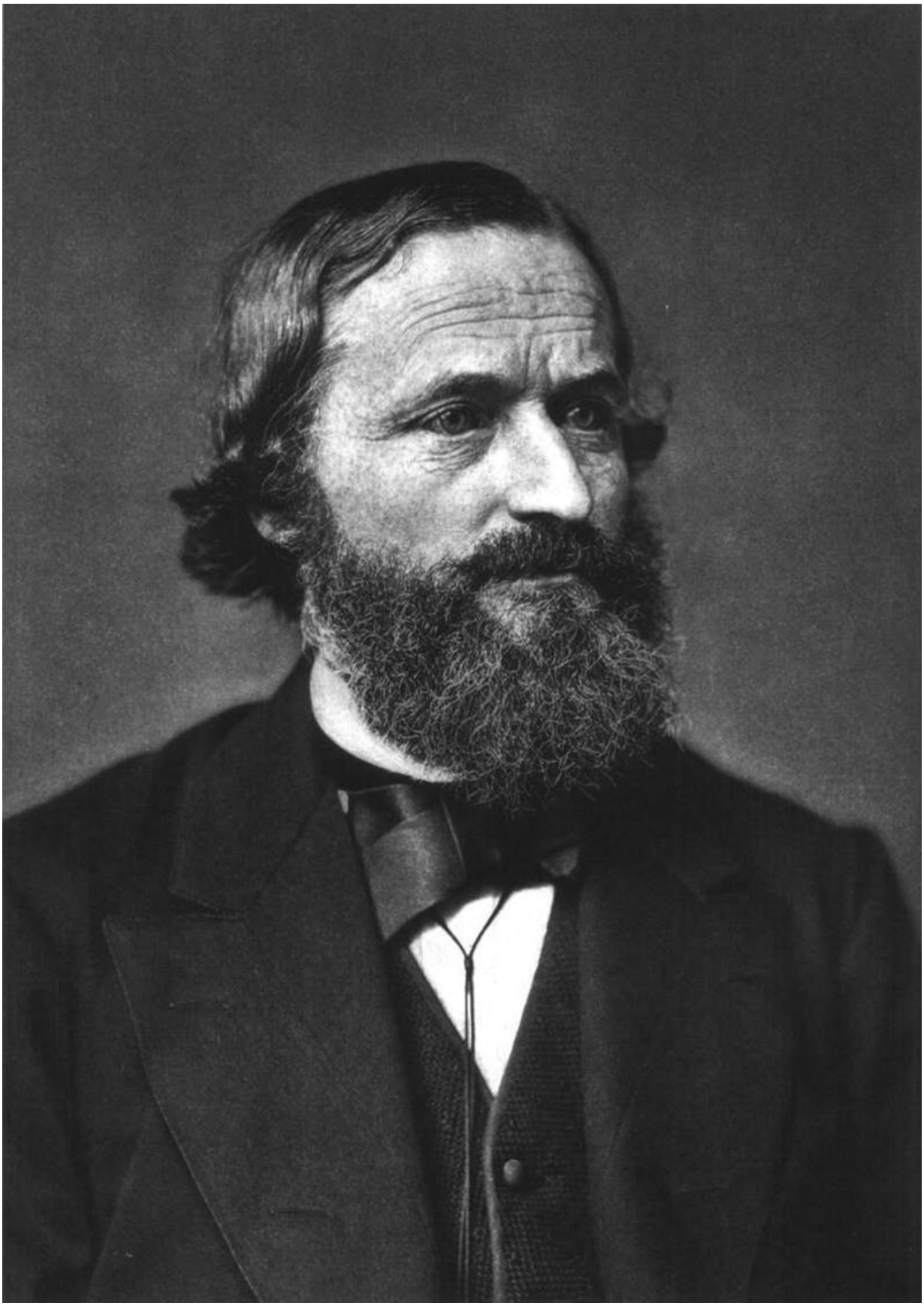


Figure 4.18 Gustav Kirchhoff defined the term "blackbody," an object which would absorb or emit all the radiation frequencies.

Source: https://en.wikipedia.org/wiki/Gustav_Kirchhoff#/media/File:Gustav_Robert_Kirchhoff.jpg.

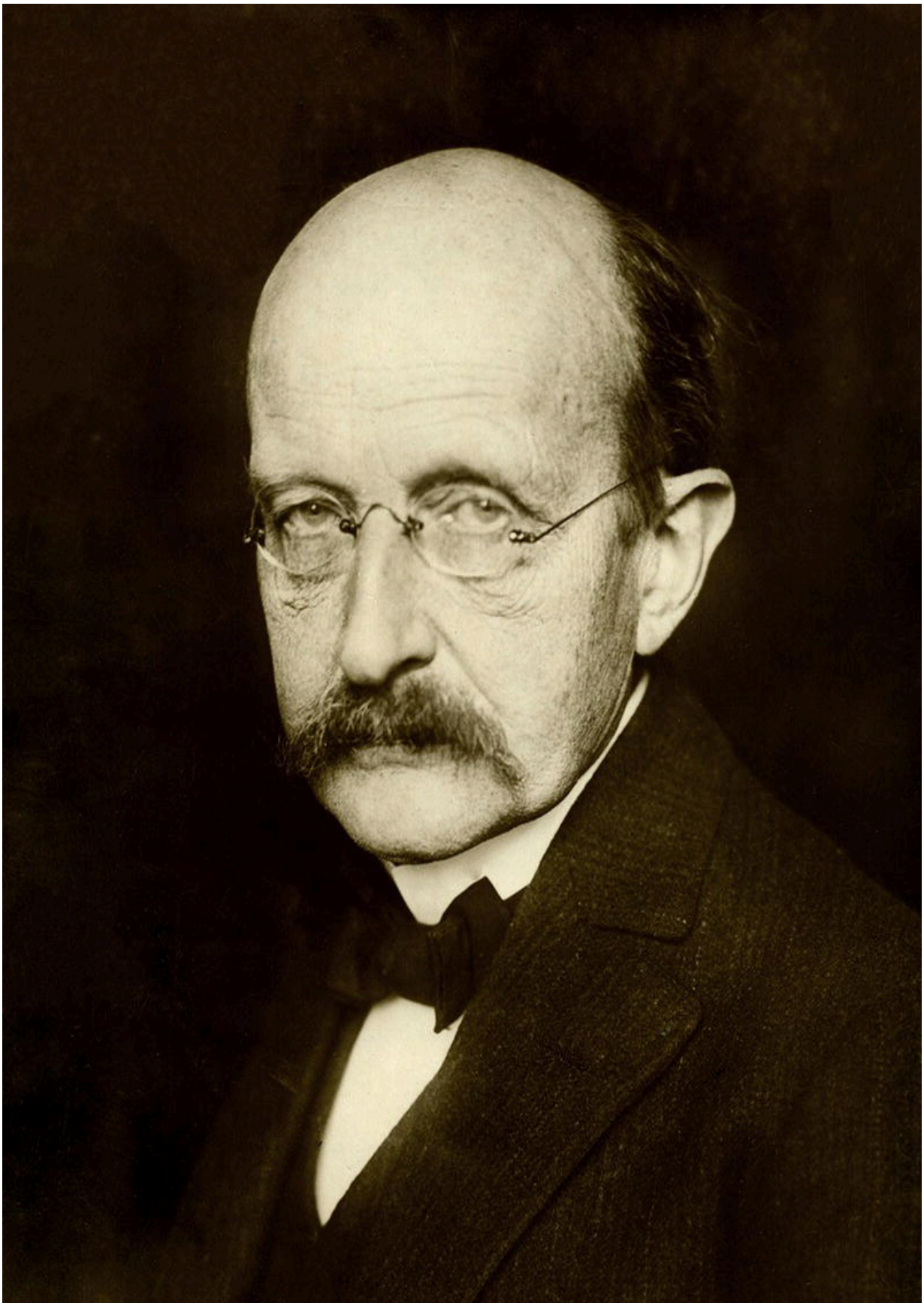


Figure 4.19 Max Planck solved the radiation problem by assuming that energies were quantized.

Source: https://en.wikipedia.org/wiki/Max_Planck#/media/File:Max_Planck_1933.jpg.

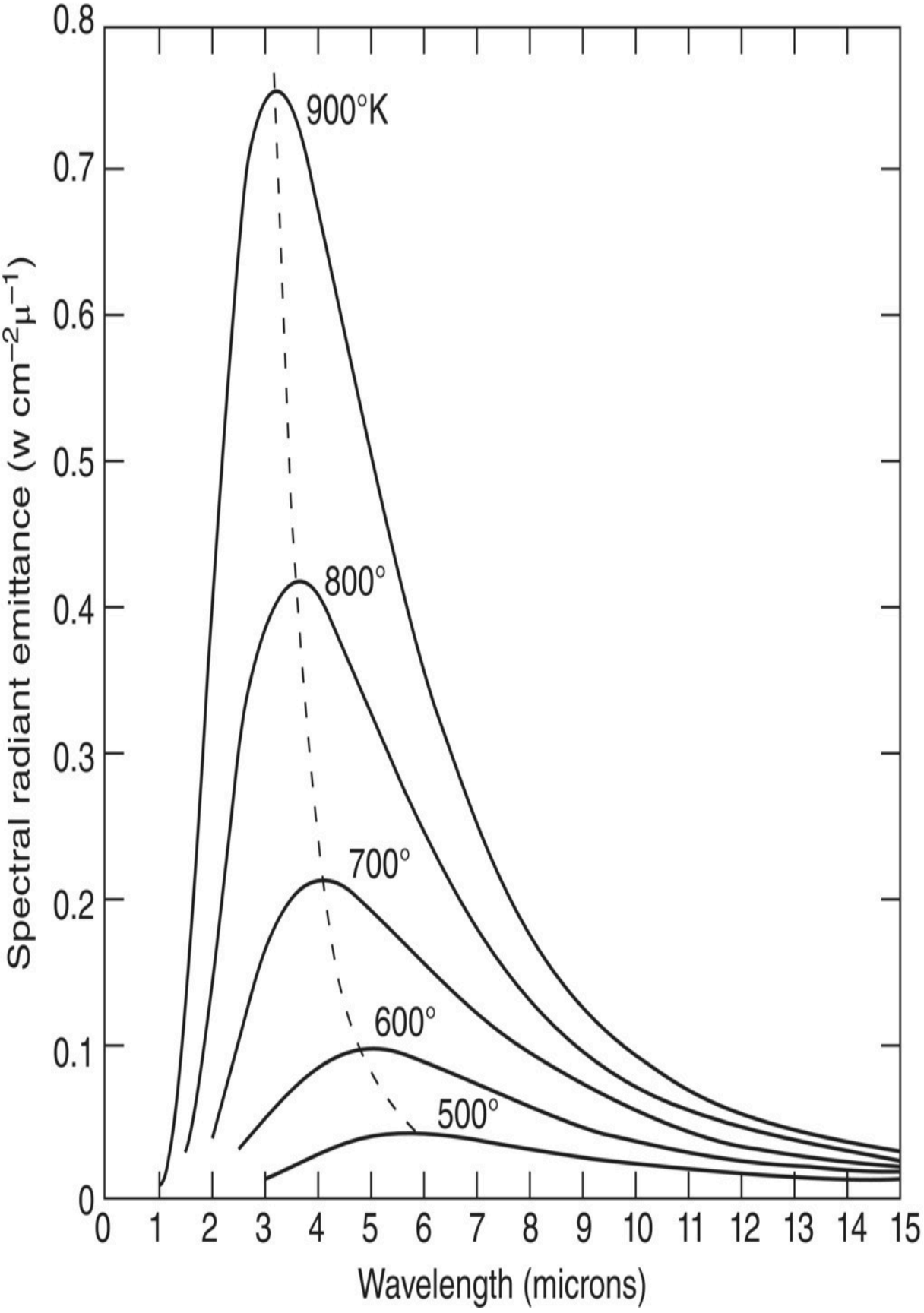

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
900°K
800°
700°
600°
500°
Spectral radiant emittance (w cm$^{-2}\mu^{-1}$)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Wavelength (microns)

Figure 4.20 The spectral emittance of a blackbody as a function of wavelength (in μm) and temperature (in K). Notice how sharp the curve gets as the temperature of the blackbody increases.

where:

W is the spectral radiant emittance, that is, how much radiation a body emits

h is Planck's constant ($6.62 \times 10^{-34}\,\mathrm{J\,s^{-1}}$)

c is the speed of light ($3 \times 10^{8}\,\mathrm{m\,s^{-1}}$)

λ is the wavelength (in μm)

k is the Boltzmann constant ($1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$)

T is the temperature (in K).

We have seen these constants before. The interesting thing about Eq. (4.7) is that there are lots of constants but only two variables, the wavelength and the temperature. So we can rewrite Eq. (4.7) as

$$W = \frac{3.7 \times 10^{4}}{\lambda^{5}} \times \frac{1}{e^{1.44 \times 10^{4}/kT}} \qquad (4.8)$$

If we plot the radiation as a function of wavelength and temperature, that is, plot Eq. (4.8), we get the curves I show in Figure 4.20.

5
The pn-Junction

OBJECTIVES OF THIS CHAPTER

After the digression in Chapter 4, talking about one of the applications that can be understood with only the knowledge of semiconductor materials, its electrical properties and the concept of energy bands and energy gaps, we are now ready to see how by combining two different semiconductor types, one p and the other n, one with extra free holes and the other with extra free electrons at room temperature, we can create very useful devices.

In this chapter I explain how a juxtaposition of one p- and one n-type semiconductor creates a pn-junction or a semiconductor diode, a device that lets the current go in just one direction. The pn-junction is the fundamental concept that explains all the semiconductor devices we use today.

5.1 The pn-Junction

Consider two boxes, one full of sand and the other empty, the situation I show at the top of Figure 5.1. What happens when I bring the two boxes together and remove any barrier between them? The sand from the left-hand box spills over to the empty box, as I show in the lower part of the figure. The sand moves to the right because of a density gradient.

Three points are very important:

First, the left-hand box has lost sand and the right-hand box has gained some.

Second, there is a force that limits and prevents more sand from sliding toward the empty box. If instead of sand I had used water, both boxes would reach the same level. This does not happen with sand because of friction, that is, there is a force that prevents the sand moving further to the right.

Third, some of the properties of the boxes have changed, for example, the left-hand box now weighs less than it did before and the right-hand box is heavier than it was when the boxes were separated.

Now consider what happens if I fabricate two separate semiconductors, one p-type and the other n-type, as I show in Figures 3.9 and 3.11. Figure 5.2 shows the conduction and valence bands of the n-type semiconductor, with the free electrons, on the left, and the p-type semiconductor with free spaces in the valence band on the right (This figure is the same as the right-hand side of Figure 3.11, except that I have added many more black balls (electrons) in the conduction band of the n-type and removed many others from the p-type semiconductor to help explain what happens when these two different doped materials are grown side by side or one on top of the other without any separation).

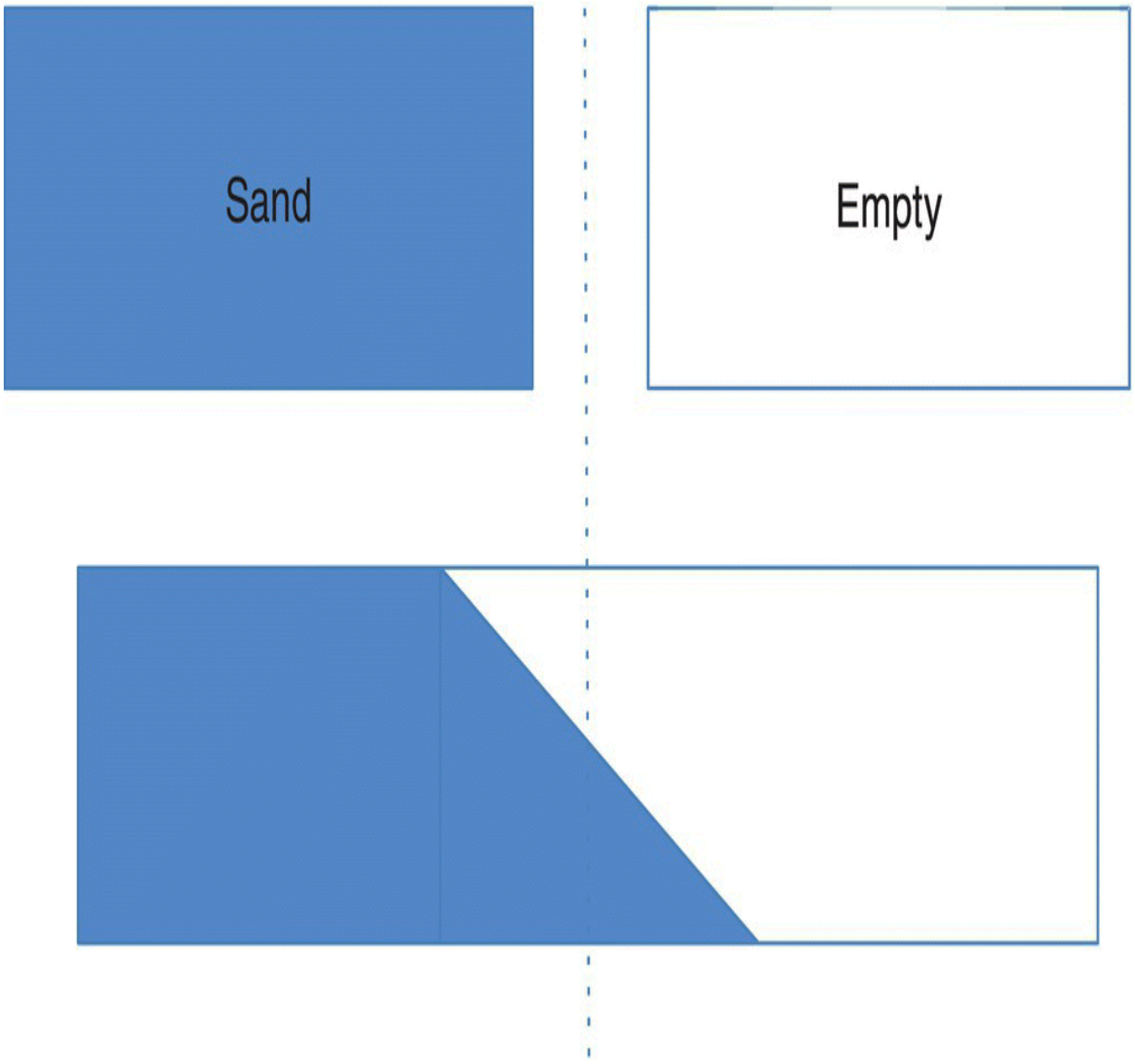


Figure 5.1 If a box full of sand is placed adjacent to an empty one, the sand spills over into the empty box on the right due to the different sand densities, eventually stopping by the counterforce of friction.

At room temperature we have seen that in an n-type semiconductor there is approximately the same number of free electrons as the number of valence 5 donor impurity atoms we have added to the silicon, between $N_D = 10^{15}$ and $N_D = 10^{18}$ impurity atoms per cm^3. Similarly, the p-type semiconductor has as many holes, empty spaces, in the valence band as the number of valence 3 acceptor

impurity atoms we have added, between $N_A = 10^{15}$ and $N_A = 10^{18}$ atoms per cm^3.

It is also very important to note that both materials, when they are separated, are electrically neutral, that is, for every free fifth electron in the conduction band of the n-type material there is one atom with five protons in the nucleus, which makes the material electrically neutral or, in other words, the electrostatic potential across the material is zero. I show this condition at the bottom of Figure 5.2. The electrostatic potential is a flat line set at zero. Similarly, with the p-type material, yes, it has holes, missing electrons, but they are compensated by the fact that they come from an element that has three protons instead of four.

Now, consider what happens when the p- and n-type semiconductors are integrally grown side by side without any separation, forming a single crystal, as shown in Figure 5.3. (we'll fabricate the junctions in Chapter 10.)

As in the above analogy of the boxes with sand (Figure 5.1) due to the different density of charges, free electrons from the n-type semiconductor move to the empty holes near the junction. We call this a *diffusion current*. Additionally, some of the electrons in the valence band of the n-type semiconductor will also move to the p-side, which is the same thing as saying that a few holes, the empty spaces, have moved from the p- to the n-type semiconductor.

The two semiconductor materials were originally neutral, but now the n-type material that has lost electrons becomes more positive (the number of protons in the nucleus of the atom has not changed) while the p-type material that has gained electrons becomes more negative. I show this electrostatic potential change at the bottom of Figure 5.3. The left-hand side is positive and the right-hand side is negative. In the middle, the potential goes smoothly from positive to negative. This is an internal electric potential that we call the *built-in potential*. The slope region between the n- and p-type regions is called the *transition region*, which is very appropriate because it

transitions from one type of semiconductor to another. It is also called the *depletion region* because this center region has lost electrons or holes to the other side, and finally it is also called the *space charged region* because there are uncompensated charges in this region (remember the atoms don't move). Lots of names for the same center region, each emphasizing a particular aspect of this middle region. Notice also that as we move away from the transition region, the semiconductors behave as when they were separated: the number of electrons and holes exactly matches the number of impurity atoms and the potential is flat.

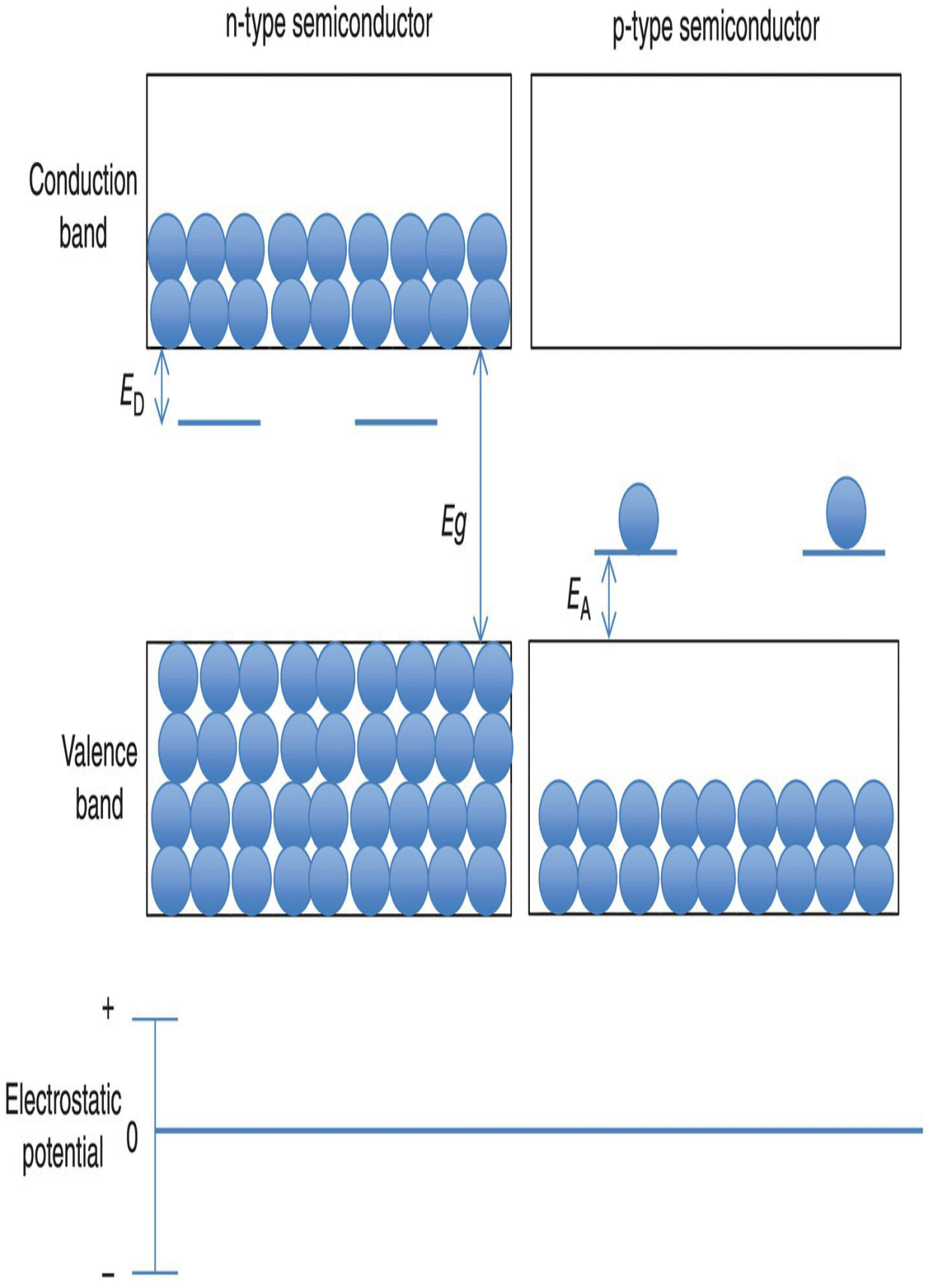

n-type semiconductor
p-type semiconductor
Conduction band
E_D
Eg
E_A
Valence band
+
Electrostatic potential
0
–

Figure 5.2 An n-type semiconductor at room temperature has lots of electrons in the conduction band and a p-type semiconductor has lots of holes in the valence band. Both semiconductors are neutral.

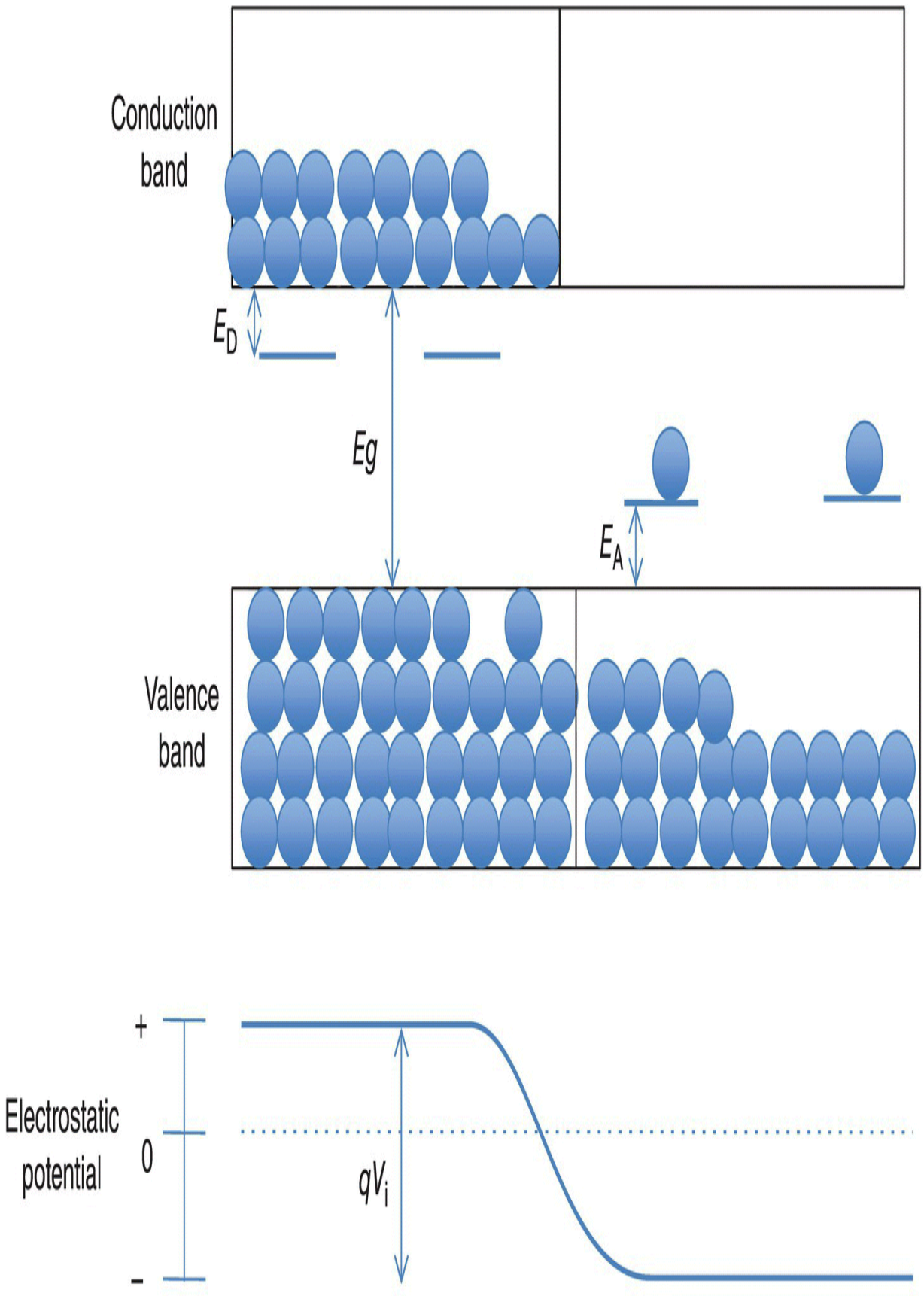

Conduction band
E_D
Eg
E_A
Valence band
+
Electrostatic potential
0
qV_i
–

Figure 5.3 When there is no separation between the p- and n-type semiconductors, free electrons from the n-type semiconductor spill over to the p-type, thus making the n-type more positive and the p-type more negative.

How many electrons move to the right by diffusion from the n- to the p-type material and how many holes move in the opposite direction? Well, notice the following: the more electrons move to the p-type material, the more positive the n-type material becomes. The electrons always want to move to the positive terminal. At some point, we reach an equilibrium condition where the diffusion force, moving the electrons from the left to the right, is equal the electric force that the electrons feel to move from the right to the left due to the electrical potential, qV_i, the drift current. I use qV_i (where q is the electronic charge) because the electrical potential is given in electron-volts. Similarly, with holes. V_i is the voltage generated by the transfer of electrons from one side to the other. This internal, intrinsic voltage, V_i, is what pushes the electrons to the right with the same strength as the different electron densities push them to the left. This forms an equilibrium condition with equal forces opposing each other. Another way of saying the same thing is to say that the *diffusion current* due to the density difference is equal to the *drift current* due to the internal electrical field. I explain the diffusion and drift current in more detail in Appendix 5.2.

Another point I like to make, because we'll use the concept later on, is that the thickness of the transition region is inversely proportional to the impurity concentration on both sides of the doped semiconductors. This makes intuitive sense. If I have few electrons in the n-type semiconductor, I will have to transfer electrons from further away into the n-region in order to generate the potential needed to stop additional flow. If instead I have a large concentration, a very thin portion of the electrons near the transition region is sufficient to create the necessary electrical field to equilibrate the drift and diffusion currents (see Appendix 5.3).

5.2 The Semiconductor Diode

Now let us see what happens when we apply an external voltage, qV_e (Figure 5.4), with the positive terminal connected to the left of the device, that is, to the n-type material.

The first thing that happens is that some of the electrons that have diffused toward the p-side are now pushed back to the n-type material because we have made the electrostatic force greater than the diffusion one. Similarly, whatever holes were in the n-type region move left toward the negative potential. But here comes the problem: for all practical purposes there are no free electrons on the p-type material so there cannot be a current moving through the circuit. Remember that to have current we need to have a continuity of charges moving in the entire loop. This condition is called the *reverse* bias voltage. After the instantaneous move of a few electrons and holes through the transition region back to their original positions, the current dies out. Only very few electrons, corresponding to the tiny intrinsic concentration, move through the transition region.

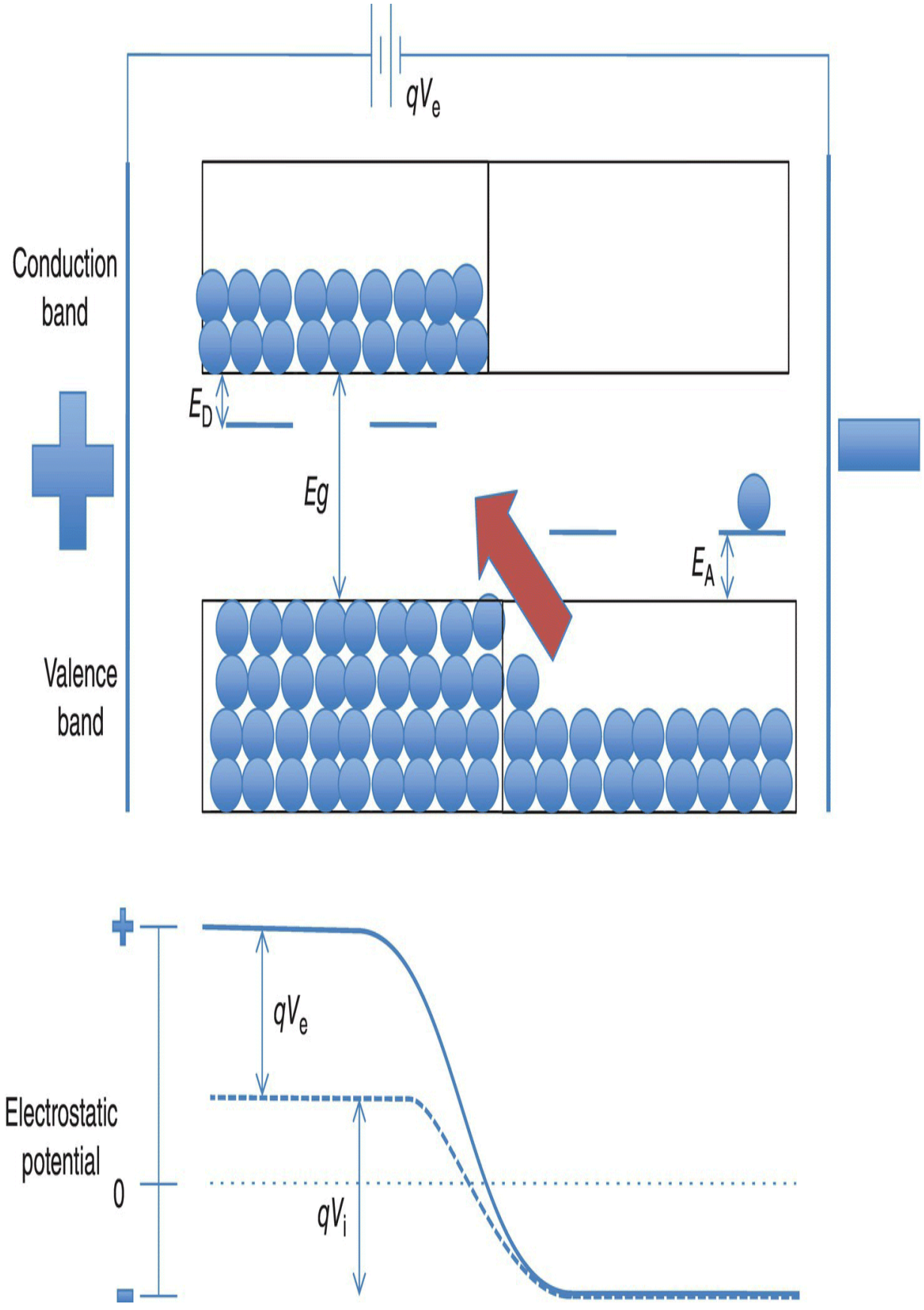
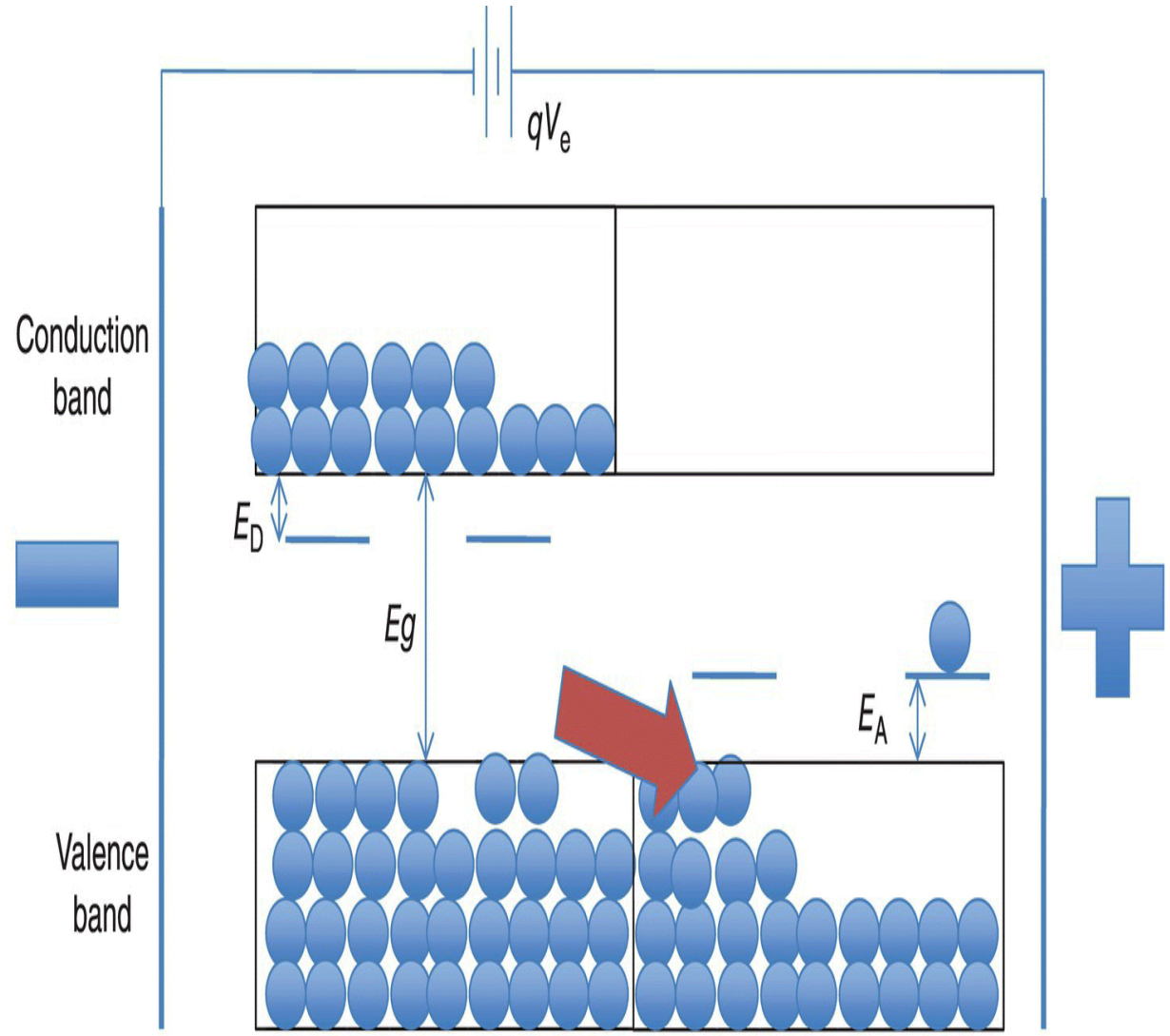

qV_e
Conduction band
E_D
Eg
E_A
Valence band
qV_e
Electrostatic potential
0
qV_i

Figure 5.4 A positive potential in the n-type semiconductor pulls electrons to the left but there are no free electrons in the p-type semiconductor and therefore no current flows. We call this the *reverse bias* condition.


qV_e
Conduction band
E_D
Eg
E_A
Valence band

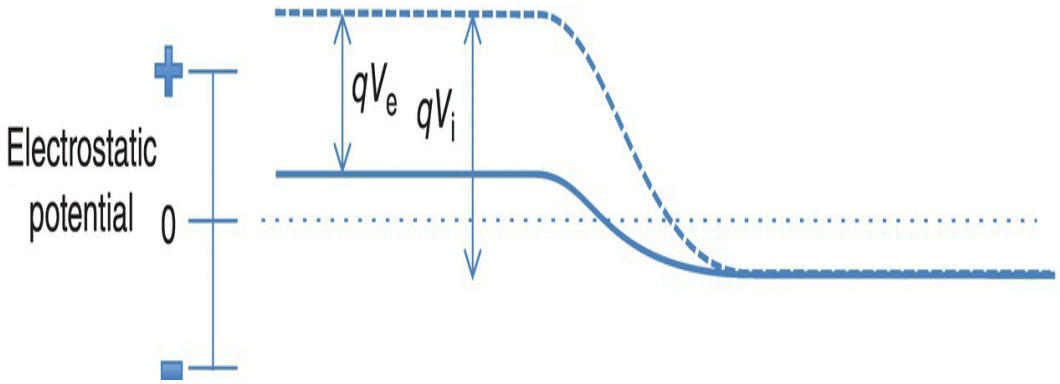

qV_e
qV_i
Electrostatic potential
0

Figure 5.5 A positive potential applied to the p-type semiconductor attracts electrons to the right and there is a large number of free electrons in the conduction band of the n-type semiconductor so current flows. We call this the *forward bias* condition.

Let's see what happens when I turn the voltage around, that is, I connect the positive terminal to the p-type semiconductor and the negative terminal to the n-type (Figure 5.5).

Now I have decreased the electrostatic barrier between the two sides but not the number of electrons and holes on either side. Therefore, there is a large number of free electrons in the n-type semiconductor ready to inundate the p-type material. Similarly, a large number of holes now move toward the negative terminal and there are lots of holes in the p-type semiconductor. As a consequence, there is a current through the pn-junction, a current that depends upon the value of the external voltage: the higher the external voltage, the higher the current. This is the condition of a *forward* biased junction.

Think of a dam. If the wall separating the water in the reservoir is higher than the level of the water in the lake, nothing changes, no matter how high the wall gets, but if I reduce the height of the wall, water will spill over.

If I plot the current in the pn-junction as a function of the applied voltage, as I increase and decrease the voltage from positive to negative the current changes, as I show in Figure 5.6.

When we forward bias the junction, that is, when we apply a positive voltage to the p-type semiconductor, with an external voltage V_e, the current increases very rapidly with increasing voltage, but if we reverse the bias, just a tiny current moves, a current limited by the very small number of free electrons in the p-type material, a number that does not change on increasing the reverse bias voltage. At very large reverse bias voltages, the pn-junction breaks down.

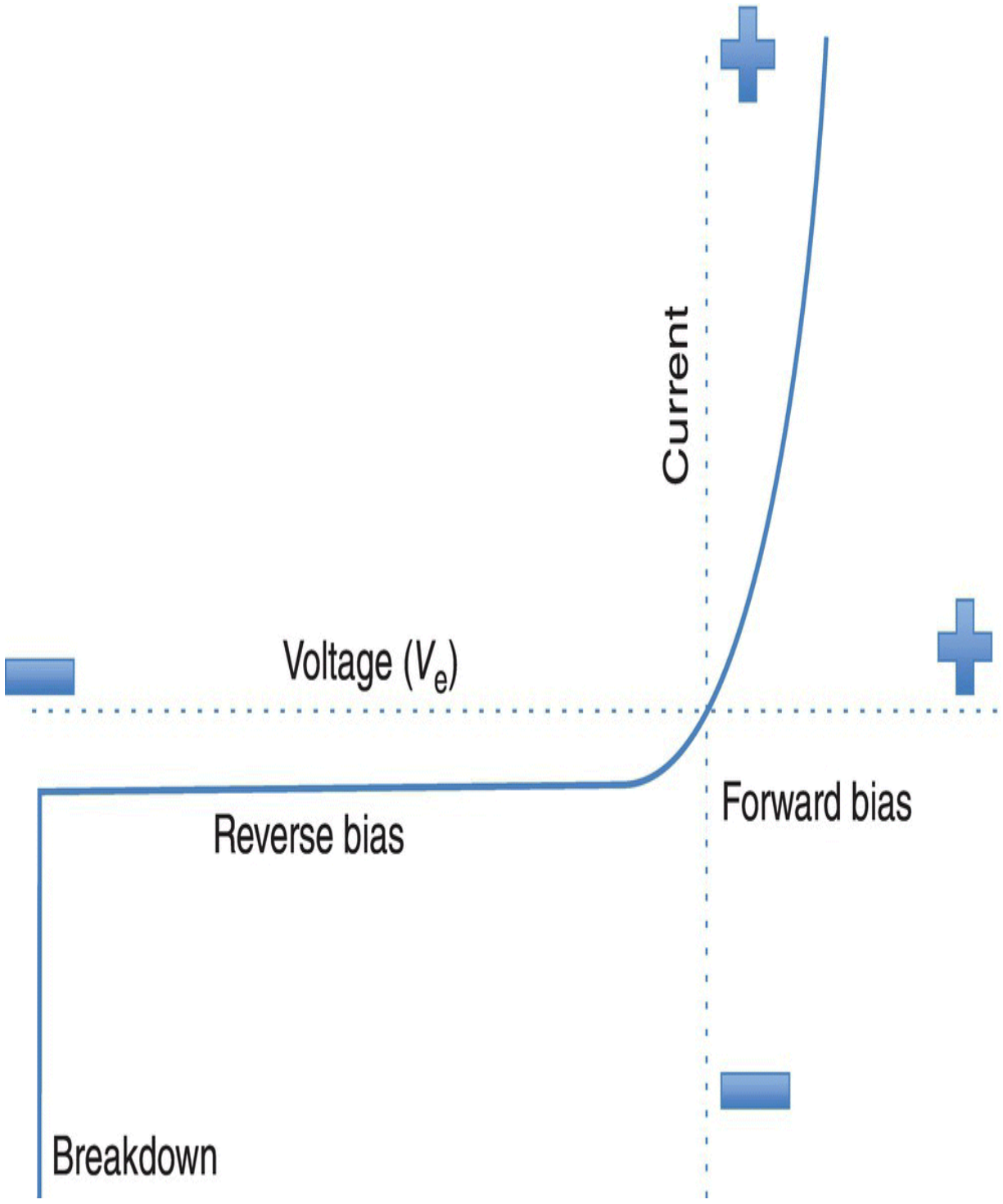


Figure 5.6 The characteristic curves of a pn-junction show current increasing when it is forward biased and practically no current when reversed biased. At some point the reversed voltage is so large that we get breakdown.

We may go back to the sand box analogy I used in Figure 5.1 but now I tilt the boxes in different directions, as shown in Figure 5.7.

Suppose that we take the boxes and first we tilt them so that the full box is lower than the empty one (left-hand side of Figure 5.7). The sand goes back to the box that contained the sand, but the upper box is empty so, the flow stops. This is analogous to the reversed biased condition. If we flip the boxes the other way, there is so much sand in the left-hand box that sand goes over to the empty box all the way to the bottom, providing the continuity of sand that we need to create a complete circuit. This is the analogous condition of a forward biased junction.

We have fabricated a diode. A diode is a device that lets the current flow in only one direction. It has a specific symbol, shown in Figure 5.8. It is a triangle indicating in which direction the positive current flows. If the positive terminal is connected to the anode, as I show in Figure 5.8, current flows. If we connect the positive side of the battery to the right, at the cathode, there is no current.

The positive charges flow from the anode to the cathode. (As an interesting point, why do we talk about positive currents when the charges moving are electrons? Supposedly, Benjamin Franklin, not knowing about the electrons, used this convention and we got stuck with it.)

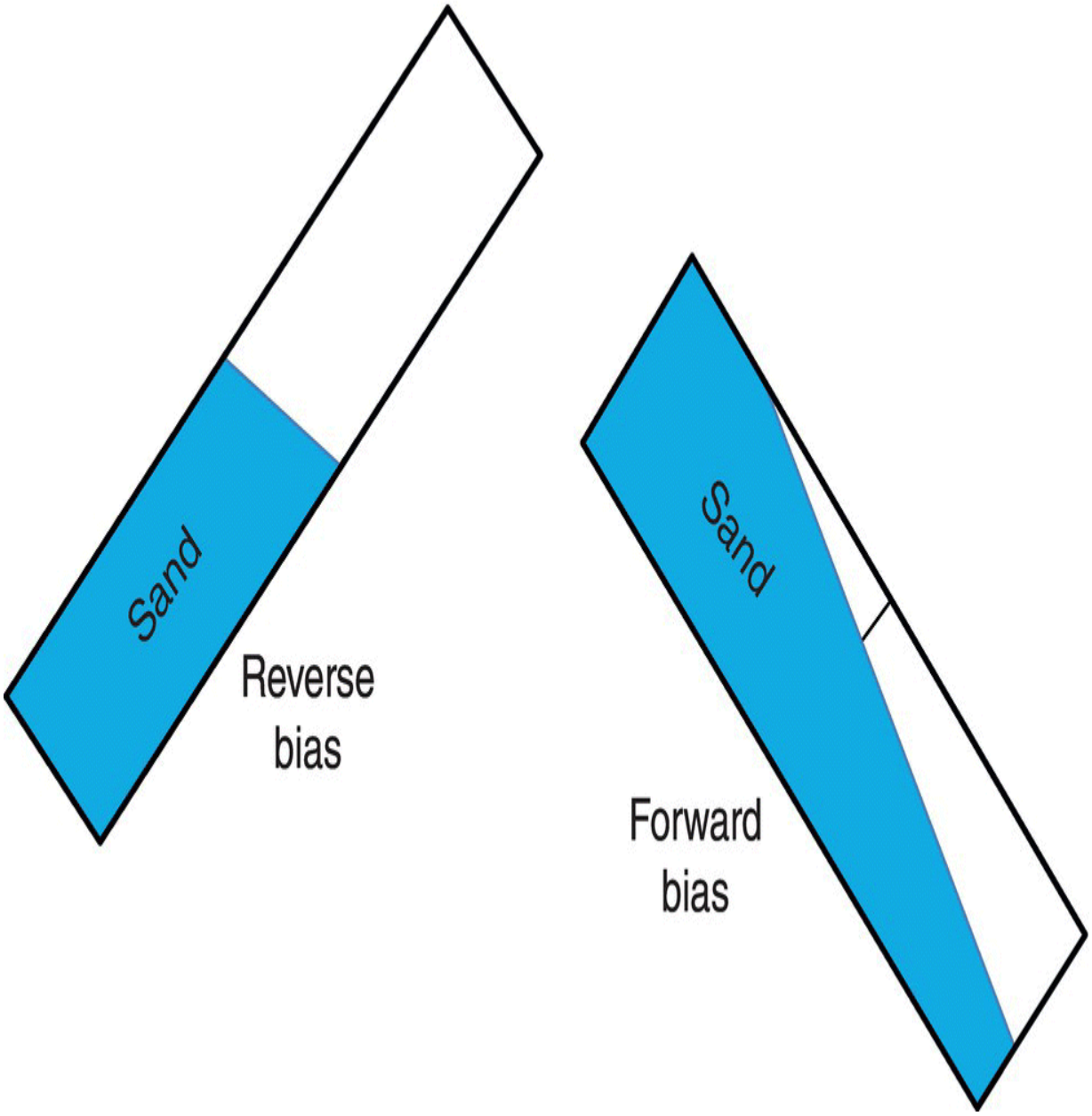


Figure 5.7 The analogy of the sand boxes with a tilt toward the full box, reversed bias, and toward the empty box, forward bias.

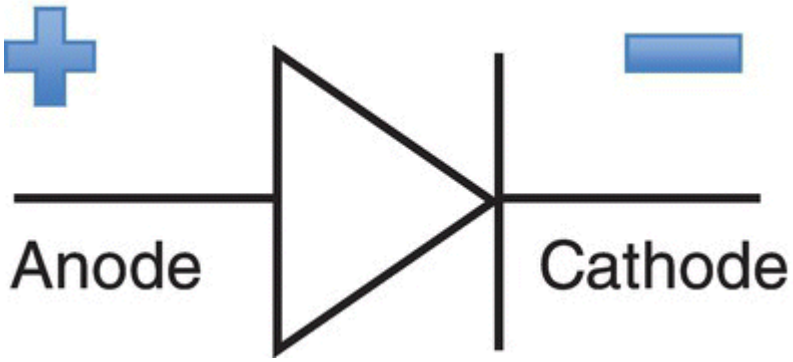


Figure 5.8 The symbol for a diode showing the direction of the current when it is forward biased from anode to cathode.

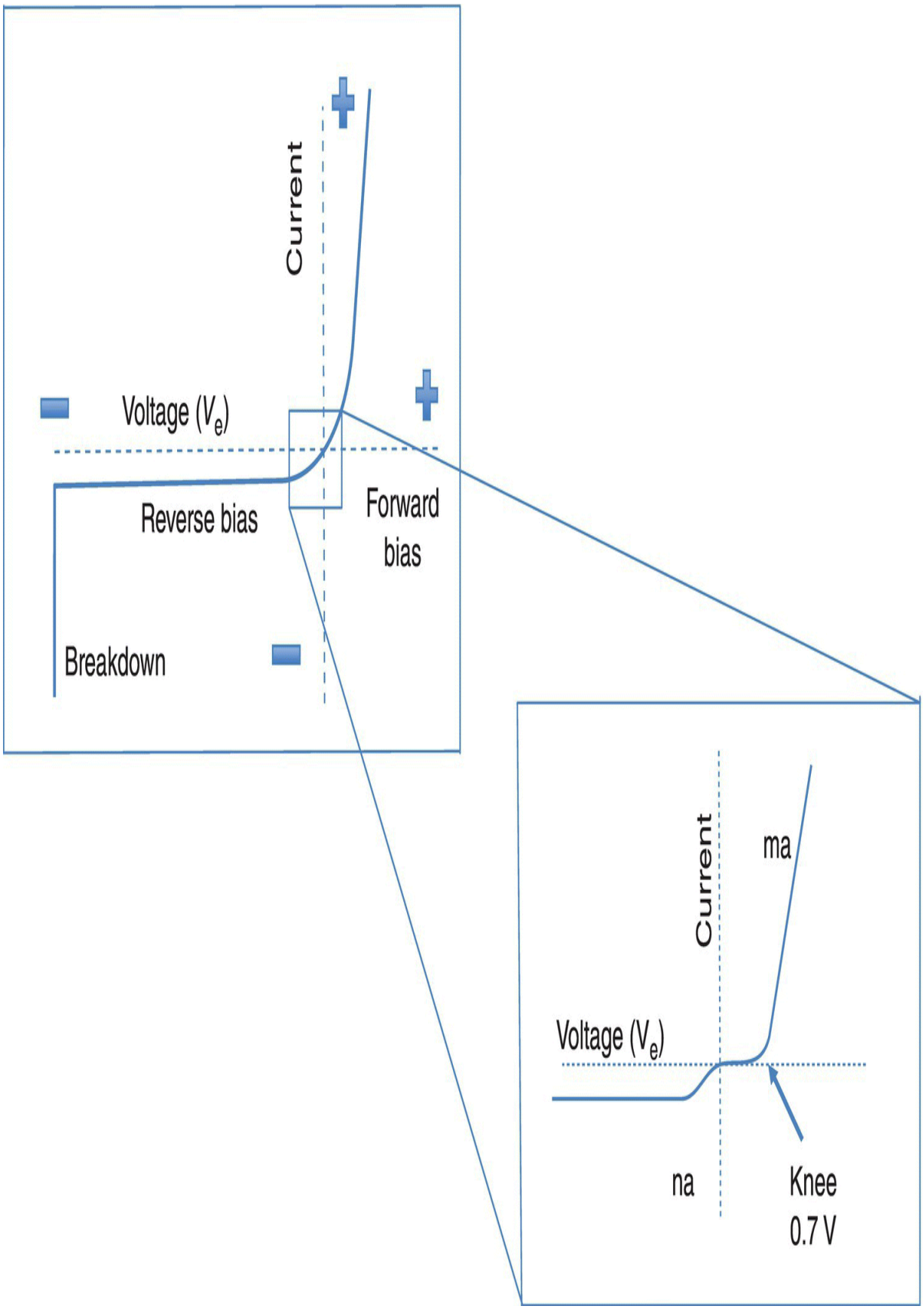

Current
Voltage (V_e)
Reverse bias
Forward bias
Breakdown
Current
ma
Voltage (V_e)
na
Knee
0.7 V

Figure 5.9 Diode characteristics showing the turn-on voltage, or the knee. Notice the change of scale: the positive current is in milliamps and the negative current in nanoamps.

The very large breakdown current I show in Figure 5.6 occurs due to the *avalanche effect*. When the reverse bias voltage is very large, one of the very few electrons in the p-type semiconductor gains sufficient energy to accelerate and hit hard an atom in the transition region, breaking a bond and creating an electron–hole pair. In turn these two charges accelerate in opposite directions, hitting other atoms, which create more electrons and holes. And the more you create, the more you accelerate, and create more and more electron–hole pairs, resulting in a very large, runaway, current.

One final comment on the diode characteristics. Figure 5.9 shows in more detail the characteristics of the diode when biased at very small voltages.

When the voltage is negative, there is a small leakage current. When I reverse the polarity, it takes some voltage before the current starts flowing. We call this the knee or the turn-on voltage. In a typical silicon semiconductor pn-junction this turn-on voltage is between 0.5 and 0.7 V.

5.3 The Schottky Diode

There are two other diodes types I like to mention, the Schottky and the Zener diodes. I show their symbols in Figure 5.10.

The Schottky diode is a diode composed of an n-type semiconductor and a metal instead of a p-type semiconductor, or vice versa, a junction of a metal to an p-type semiconductor. The main advantage of the Schottky diode is that its turn-on voltage (see Figure 5.9) is much lower than that of the semiconductor diode, 0.2 V versus 0.7 V. This allows the diode to function at much faster switching speeds. Because the Schottky diode turns on much sooner at a lower voltage, there is less heat dissipated, which makes it very important in digital microcircuits where time and heat are real concerns.

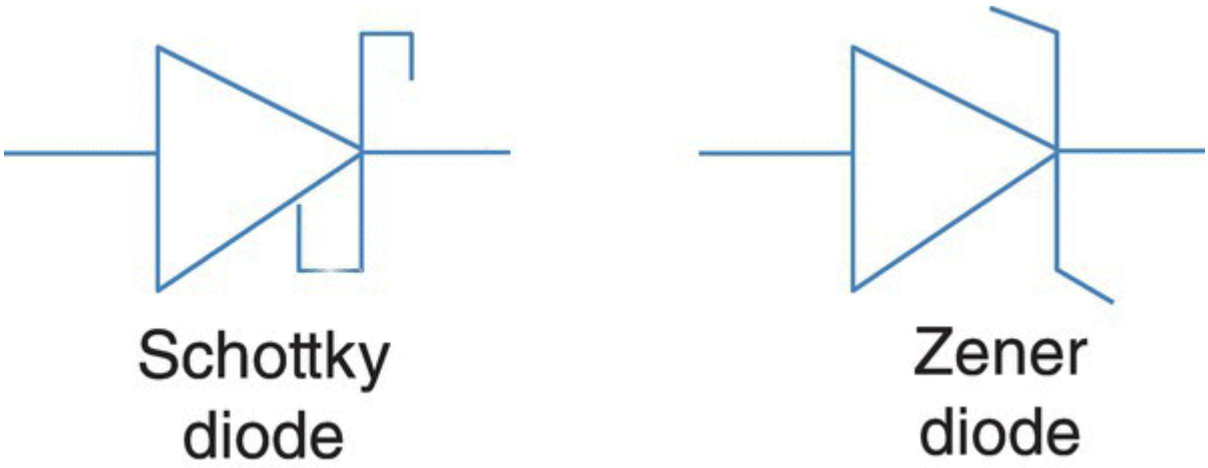


Figure 5.10 Symbols for Schottky and Zener diodes.

The mechanism of the Schottky diode is slightly different to that of a pn-junction. To transfer an electron from a metal to a semiconductor or vice versa, the electron has to leave one material to enter another. The energy needed by an electron to escape a solid is called the *work function*, *W*, which is the energy difference between where the electrons are in the solid and the energy they need to escape (I explain this in more detail in Appendix 5.4). The electrons don't care how many electrons are in one material or the other. They just move to whatever material has lower energy levels.

Suppose, for example, that you have two containers side by side, as shown in Figure 5.11.

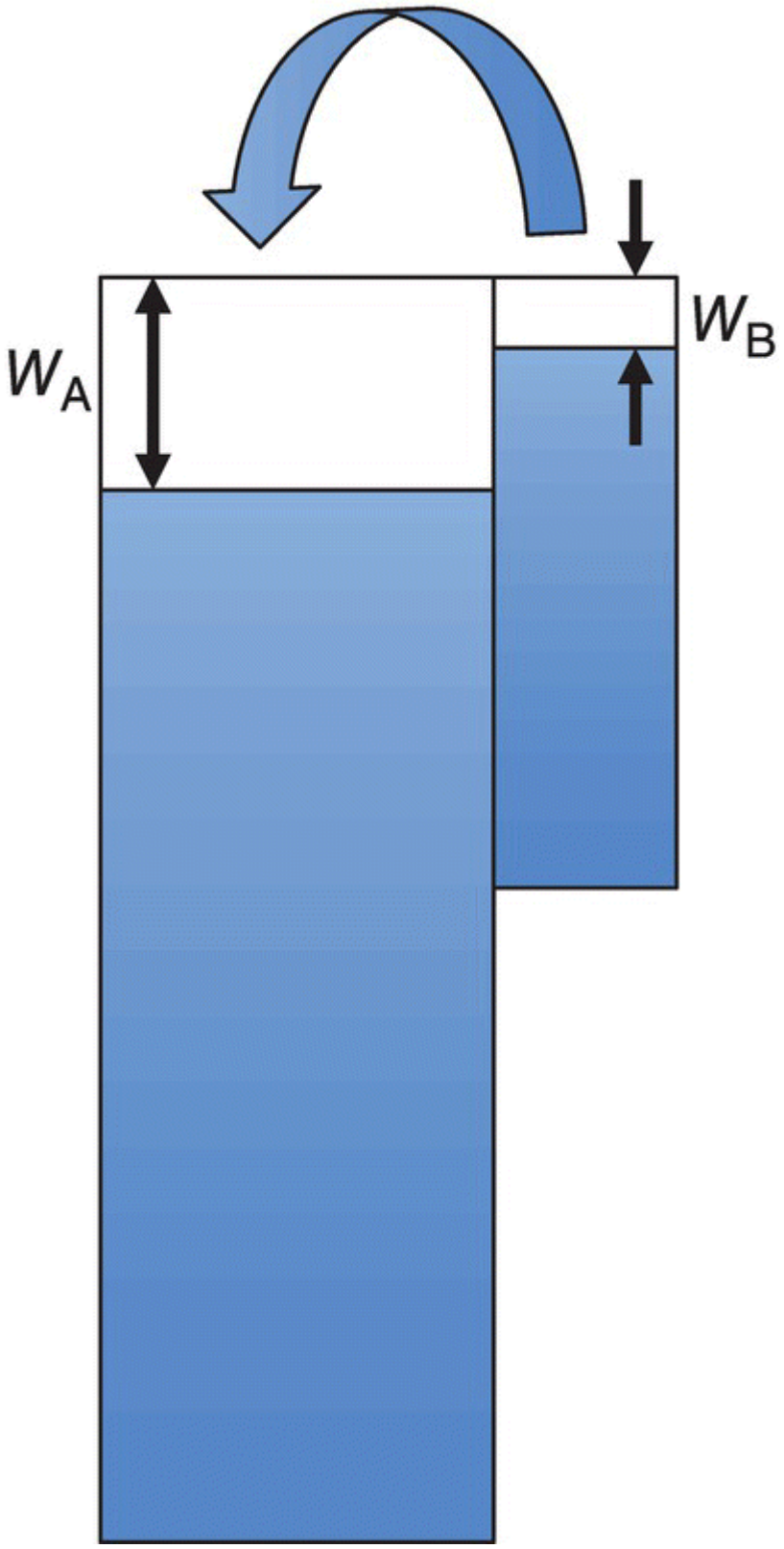

W_A
W_B

Figure 5.11 The water in the small container on the right will boil over into the one on the left because W_B is less than W_A.

The container at the left is very large with lots of water. The one on the right is much smaller and contains much less water. I set them up so the rims of the containers are at the same level. If I apply heat to both containers and the water starts boiling, the water will move from the small to the large container because the level of the water in the large container relative to the rim of the container is lower than that on the right. This is basically what happens in the Schottky diode. The electrons move from the n-type semiconductor to the metal because the work function of the n-type semiconductor is lower than that of the metal.

Another advantage of Schottky diodes is that they can switch much faster, almost instantaneously, because the junction is so thin that there is no need to move electrons and holes around as in the larger transition regions, see Appendix 5.3.

5.4 The Zener or Tunnel Diode

The Zener diode, whose symbol I show on the right in Figure 5.10, is a diode that allows current to flow under reverse bias conditions. The Zener diode is a device based on truly quantum mechanical concepts. Classical physics cannot explain its operation.

Let me start by saying that in quantum mechanics the electrons are both particles and waves, or maybe better, an electron can behave as a particle or as a wave. In a classical system (Figure 5.12) if I throw a ball against a barrier, the ball slows down as it rise up the barrier and it will go over the barrier if and only if the speed of the ball, its kinetic energy, is larger than the potential energy that the ball would have at the top of the barrier. If not, in classical physics, the ball stops before it reaches the top and goes backwards, moving in the opposite direction with the same speed as it had coming up (assuming, of course, frictionless surfaces). This is true no matter how thick or thin the barrier is. The ball goes over the barrier only

and exclusively depending on how high the barrier is, not its thickness.

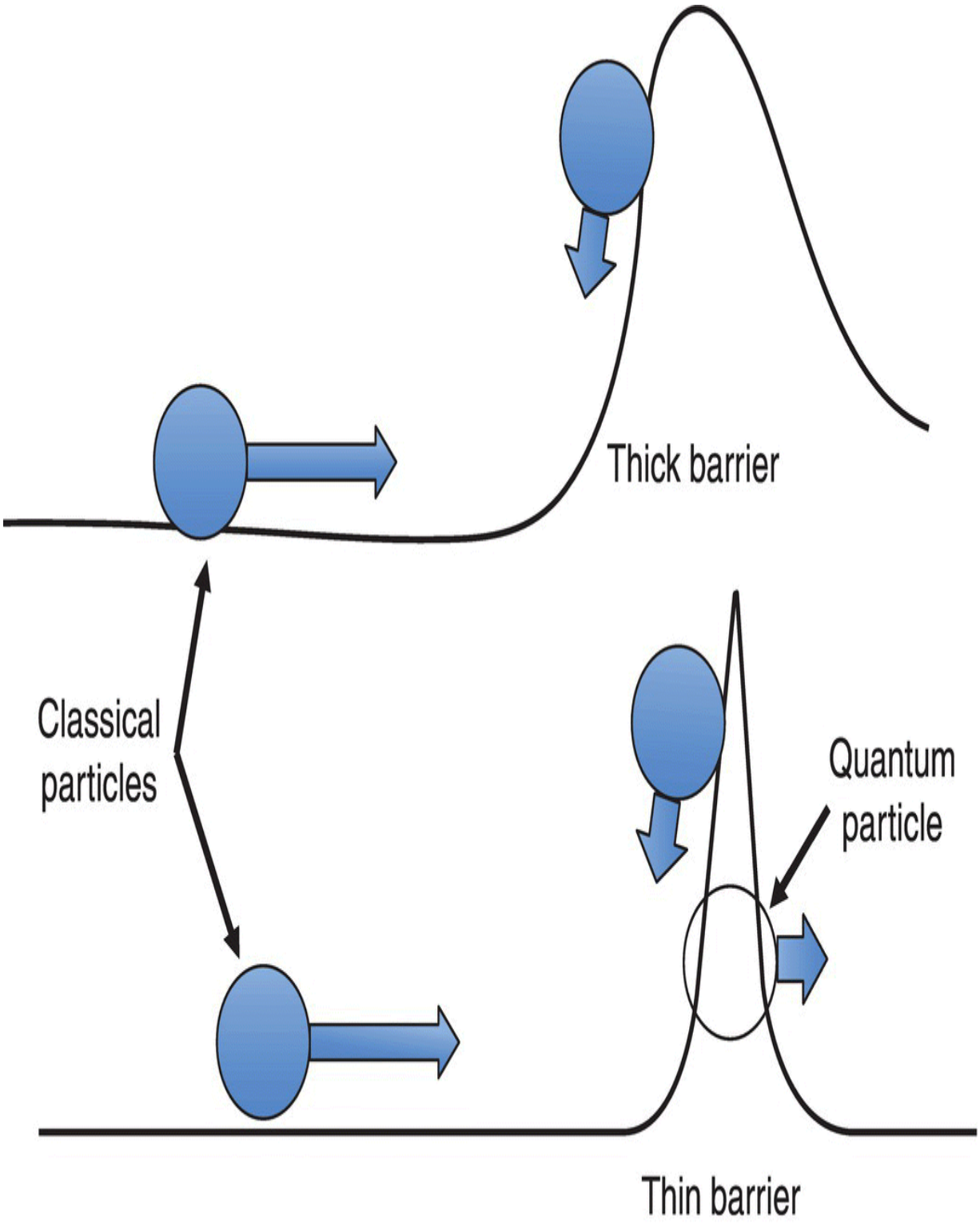

Thick barrier
Classical
particles
Quantum
particle
Thin barrier

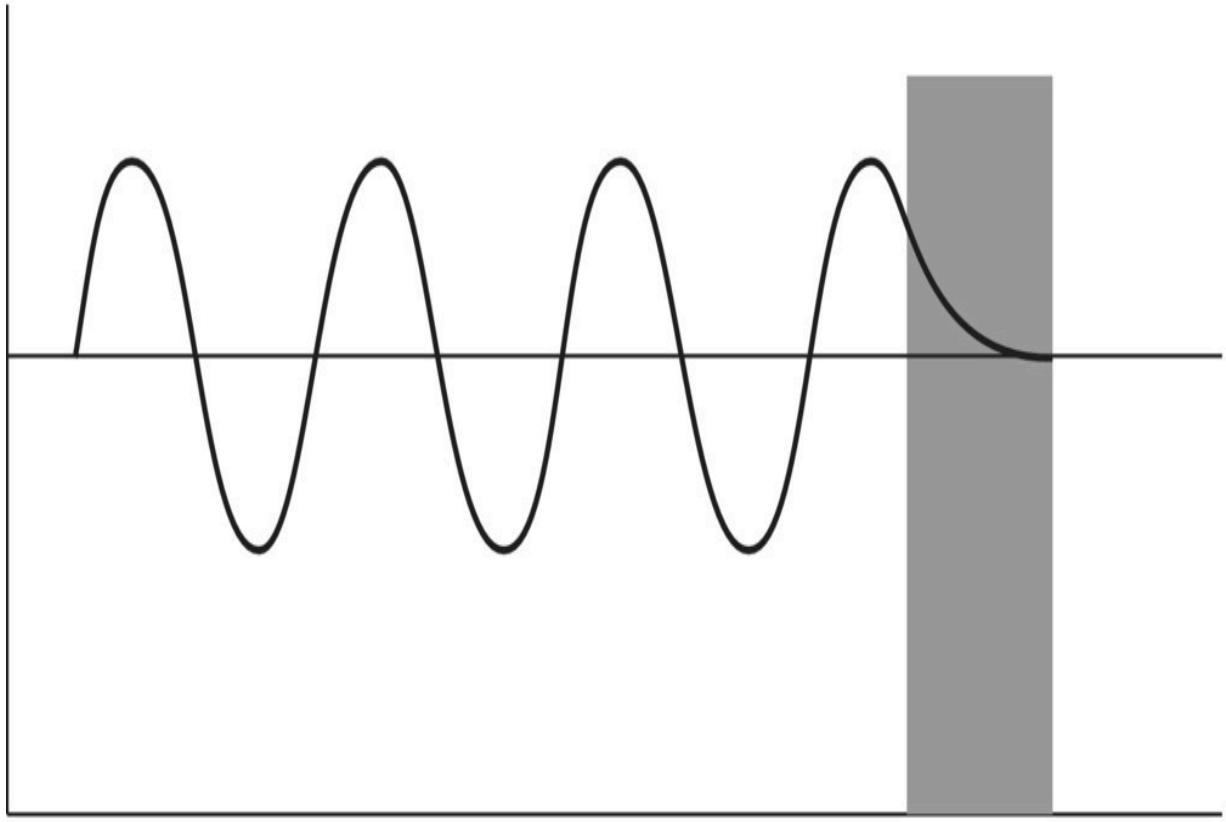
Figure 5.12 A classical ball will cross the barrier only if its energy is high, but quantum mechanics tells us that an electron can penetrate the barrier depending on not only how high the barrier is, but also how thin.

In the case of a quantum wave, the situation is different. There is a probability that the quantum particle, an electron for example, goes through the wall, that is, tunnels through, if the wall is thin enough. Figure 5.13 represents the electron as a wave that shows the probability that the particle is found anywhere, including at the other side of the barrier.

As the wave hits a thick potential barrier, its amplitude quickly decreases as I show in the upper sketch of Figure 5.13. The taller or the wider the potential barrier is, the faster the wave decays to nothing. Thus, the probability of finding the electron at the other side is zero. If I make the potential barrier very thin (lower drawing), the amplitude of the wave decreases but the wave, with decayed amplitude, crosses the barrier and appears at the other side. In the classical case the ball never crosses the high barrier, but in quantum mechanics, the wave crosses the barrier depending on both its height and its thickness. The quantum mechanical wave tells us the probability of where to find the electron. Thus, the situation shown in the lower part of Figure 5.13 tells me that there is a probability that I will find the electron at the other side of the barrier, even though the potential is higher than the energy of the electron. I should also mention that the wave is not the electron; the wave just tells us the probability of where the electron is or will be. Don't try to visualize that; it is quantum mechanics after all.

After this extremely brief introduction to quantum physics, let me explain the Zener diode, also known as the tunnel diode. If we have very highly doped p- and n-type materials forming the junction (between 10^{18} and 10^{19} impurities/cm^3), the transition region becomes high and very narrow (see Appendix 5.3). The energy bands, without any outside voltage, look like those in A in Figure 5.14. The transition region is very narrow and electrons in the

valence band of the p-type semiconductor are facing the electrons in the conduction band of the n-type semiconductors. Nothing happens. But suppose I reverse bias the diode, as shown in part B of Figure 5.14. Now a large number of electrons from the p-type semiconductor are separated by a very thin barrier from a large portion of free allowed energy spaces in the conduction band of the n-type semiconductor and they are able to tunnel through, therefore a reversed current flows from the p to the n side. Take a look now at Figure 5.15. Without any bias, at point A in Figure 5.15 the current is zero. When we apply a negative voltage there is a large negative current tunneling through the diode, point B in Figure 5.15. Now back to C in Figure 5.14. We apply a very small forward bias voltage. The electrons in the conduction band of the n-type material are facing the empty energy levels in the p-type material also across a very narrow gap. Therefore, a current flows from the n- to the p-type semiconductors. I show this current in Figure 5.15 in area C, between 0 and 0.1 V. As I increase the forward bias further, D in Figure 5.14, the separation between the electrons in the n-type material and the holes in the p-type material gets larger, the tunneling gets harder, and the current starts decreasing as I show in Figure 5.15, region D, between 0.1 and 0.2 V. Finally, as we keep increasing the forward bias, the Zener diode starts behaving like a regular forward bias diode and the current starts increasing quickly, as it does in the regular diode (Figure 5.6). I show this in region E of Figure 5.15.



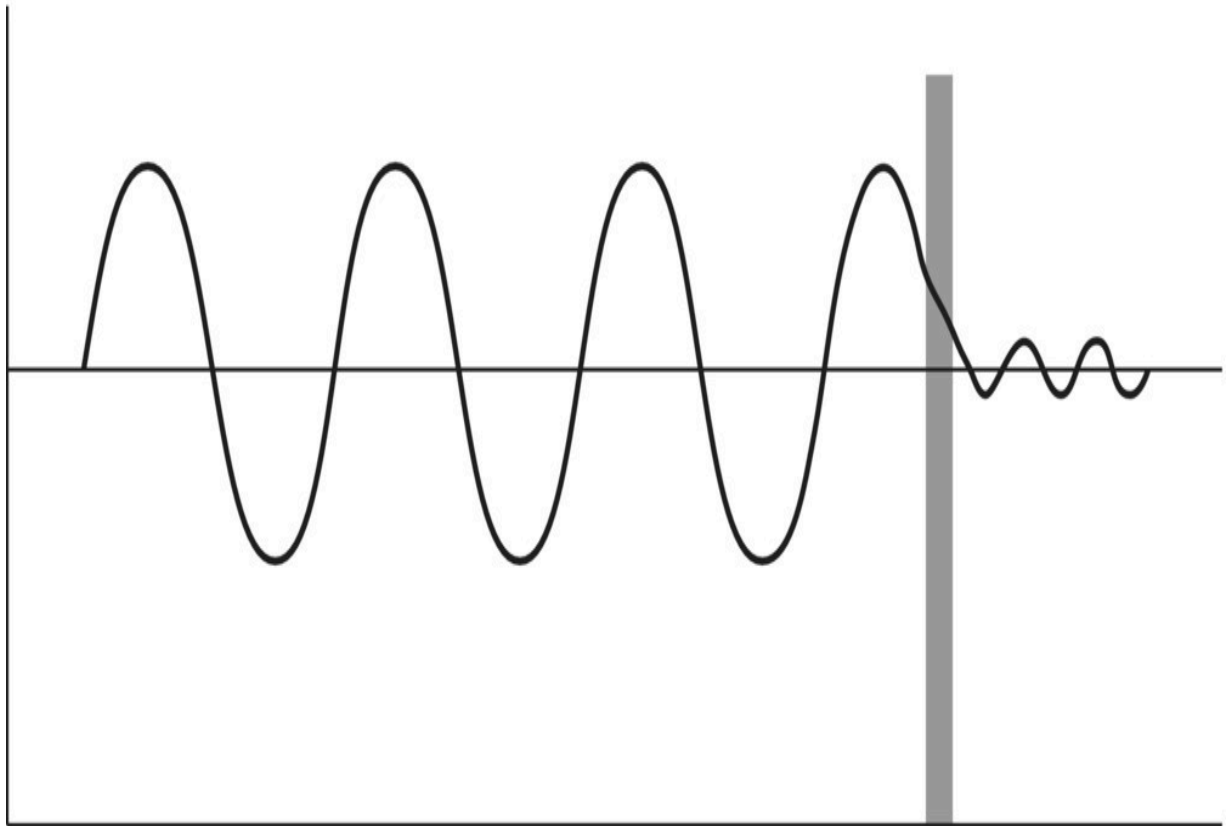


Figure 5.13 In quantum mechanics the probability of finding an electron is expressed by its wave function. If the barrier is thin enough, there is a probability that the electron will be found on the other side of the barrier.

Region D is interesting. It shows that as the voltage increases from 0.1 to 0.2, the current actually decreases. This is a region of negative resistance. Since the tunneling is almost instantaneous, tunnel diodes are used in high-speed devices in the gigahertz region and for ultra-fast switches. The negative resistance is also very useful for designing electronic oscillators. Reverse bias currents are very high so one needs to be careful not to burn the device.

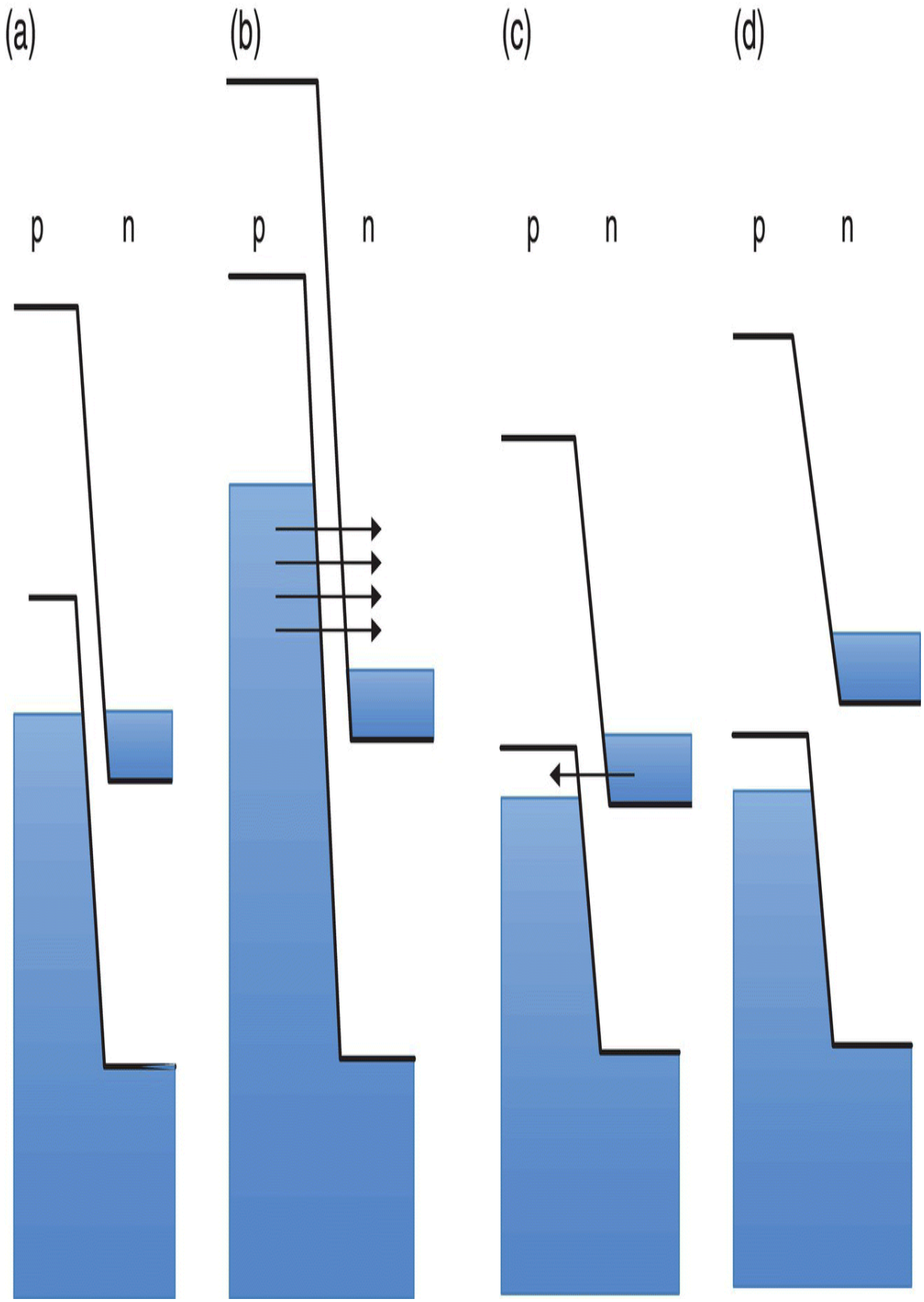

(a)
(b)
(c)
(d)
p
n
p
n
p
n
p
n

Figure 5.14 A Zener diode has such a thin transition region (A), that electrons can cross the barrier under reverse bias conditions (B) and under a small forward voltage (C). Further increase of forward voltage increases the barrier distance (D) stopping the tunneling current through it.

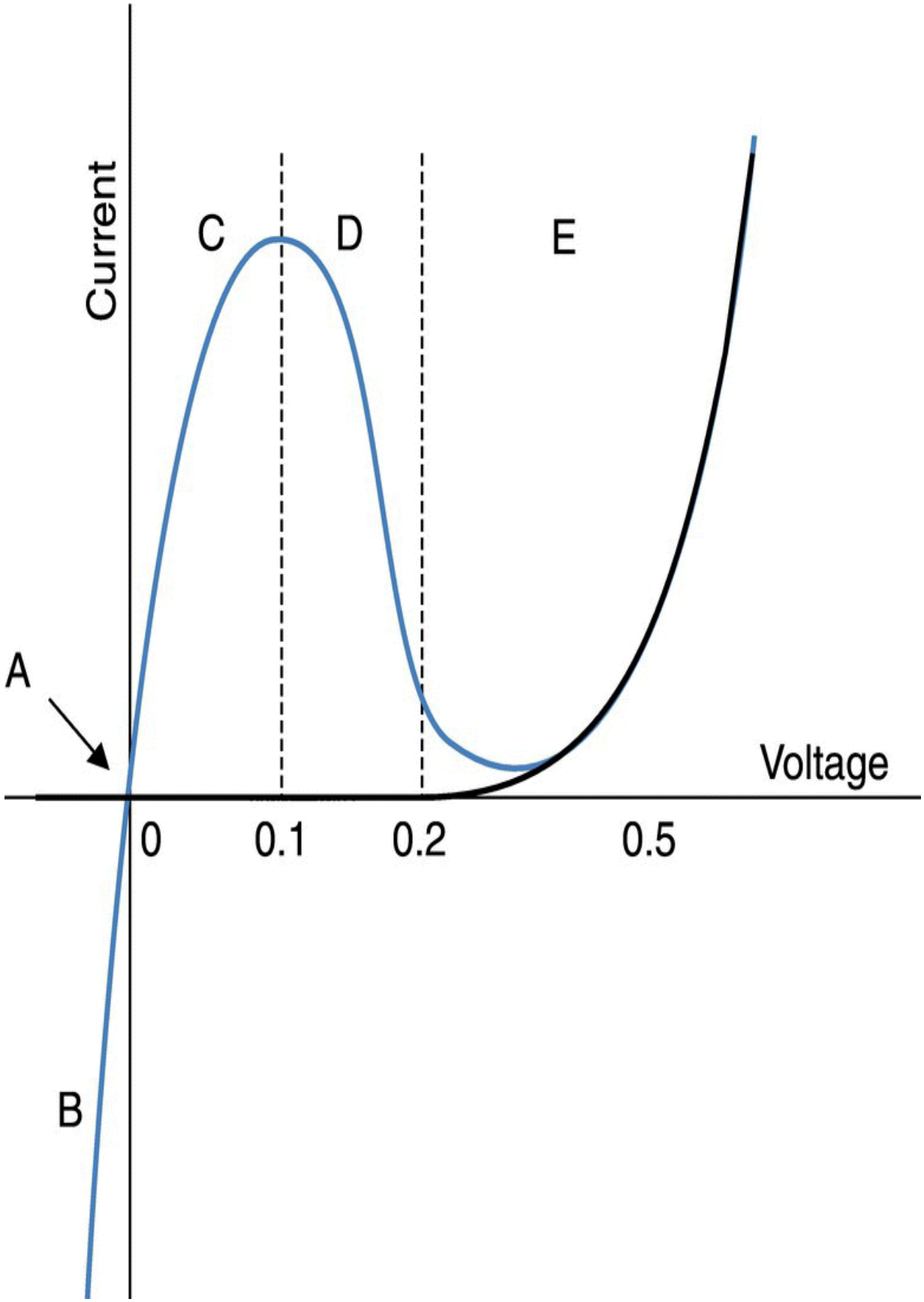

Current
C
D
E
A
Voltage
0
0.1
0.2
0.5
B

Figure 5.15 The tunnel diode characteristics show a high reverse bias current (B) and a negative resistance, region D.

5.5 Summary and Conclusions

We have seen how a combination of two different types of semiconductors, p and n, create a pn-junction and generate a transition region which creates an internal electrical voltage. In equilibrium, the diffusion current due to the difference of electrons and holes density cancels the drift current due to this intrinsic electric field so the current outside the diode is zero (Figure 5.3). Forward biasing the diode (positive terminal at the p-semiconductor) results in large currents and reversing the bias (positive terminal at the n-semiconductor) results in no current except for a very small leakage current (Figure 5.6).

We have also discussed two variations of the diode. The Schottky diode replaces one of the semiconductors by a metal, making it a faster and lower power diode. The tunnel diode, which can be explained only in terms of quantum mechanics, has a large reverse bias current and also results in a region of negative resistance.

I urge you to read or at least look at the appendices. Some of the details will help you to understand and hopefully even clarify the behavior of pn-junctions and metal–semiconductor junctions.

In Chapter 7, after a digression explaining what we call "passive element" resistors, capacitors, and inductors in Chapter 6, I go over many of the applications of these semiconductor diodes.

Appendix 5.1 Fermi Levels of a pn-Junction

I introduced you to the Fermi level in the appendices of Chapters 2 and 3. Let's see what happens when we apply the same Fermi–Dirac (F-D) statistics to pn-junctions. Figure 5.16 shows the situation where the pn-junction is at 0 K.

In a system, the Fermi level must be the same throughout. The Fermi level can be visualized as the water level in connected vessels of different sizes and heights. The level of the water in a lake does not change from the shore to the deepest location. Figure 5.16 is the same as Figure 3.16, except that I have aligned the dotted lines showing the location of the Fermi level in the n- and p-type semiconductors. At 0 K, all the lowest allowed energy sites in the p, n, and transitions regions are occupied with electrons and all those above the Fermi level are empty. Therefore, on the n-type side, the electrons are all in the valence band or in the energy levels of the donor band. In the p-type semiconductor all the electrons are in the valence band, and the acceptor levels, even though they are very close to the valence band, are empty. The valence and conduction bands have to somehow transition smoothly from the n- to the p-sides.

Now consider what happens when we have the pn-junction sitting on the table at room temperature, 300 K (Figure 5.17). At room temperature, the F-D function, on the left of the figure (the same identical F-D function I used in Appendices 2.2 and 3.1) centered at the Fermi level, predicts that the electrons in the donor band move to the conduction band, and the empty levels of the acceptor bands are occupied by electrons, leaving a large number of holes in the valence band of the p-semiconductor. This is exactly the same as I mentioned above and I showed in Figure 5.3, but now the Fermi function confirms and lets us calculate numerically the number of free electrons in the conduction band and free holes in the valence band. It also lets us calculate the number of electrons and holes in the transition region. (Figure 5.3 show many more electrons than Figure 5.16. They are the same, but I use different graphics to better understand the concepts. I hope this is not confusing.)

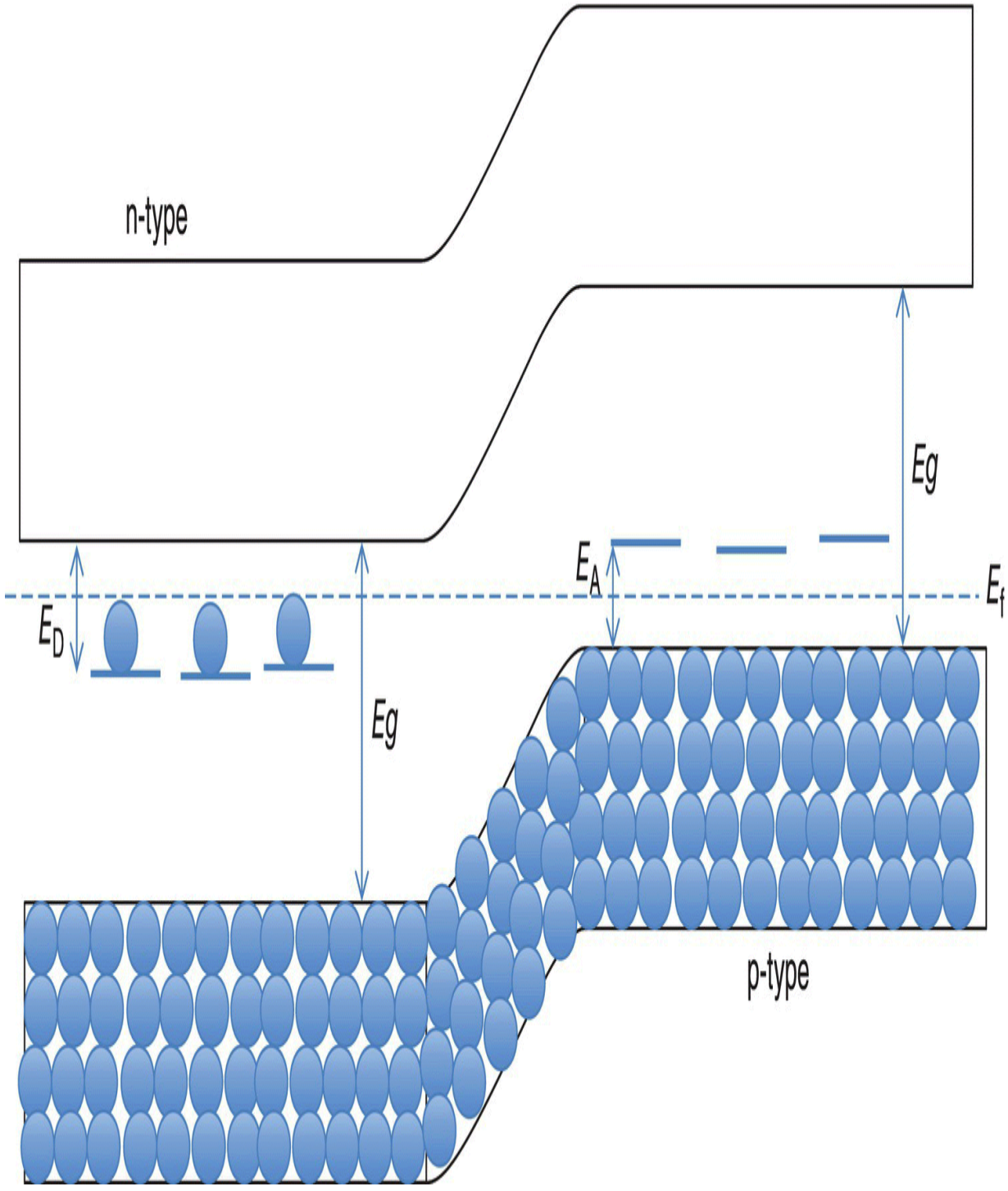


Figure 5.16 The pn-junction at 0 K has all the levels below the Fermi level occupied with electrons and all the energy sites above the Fermi level are empty.

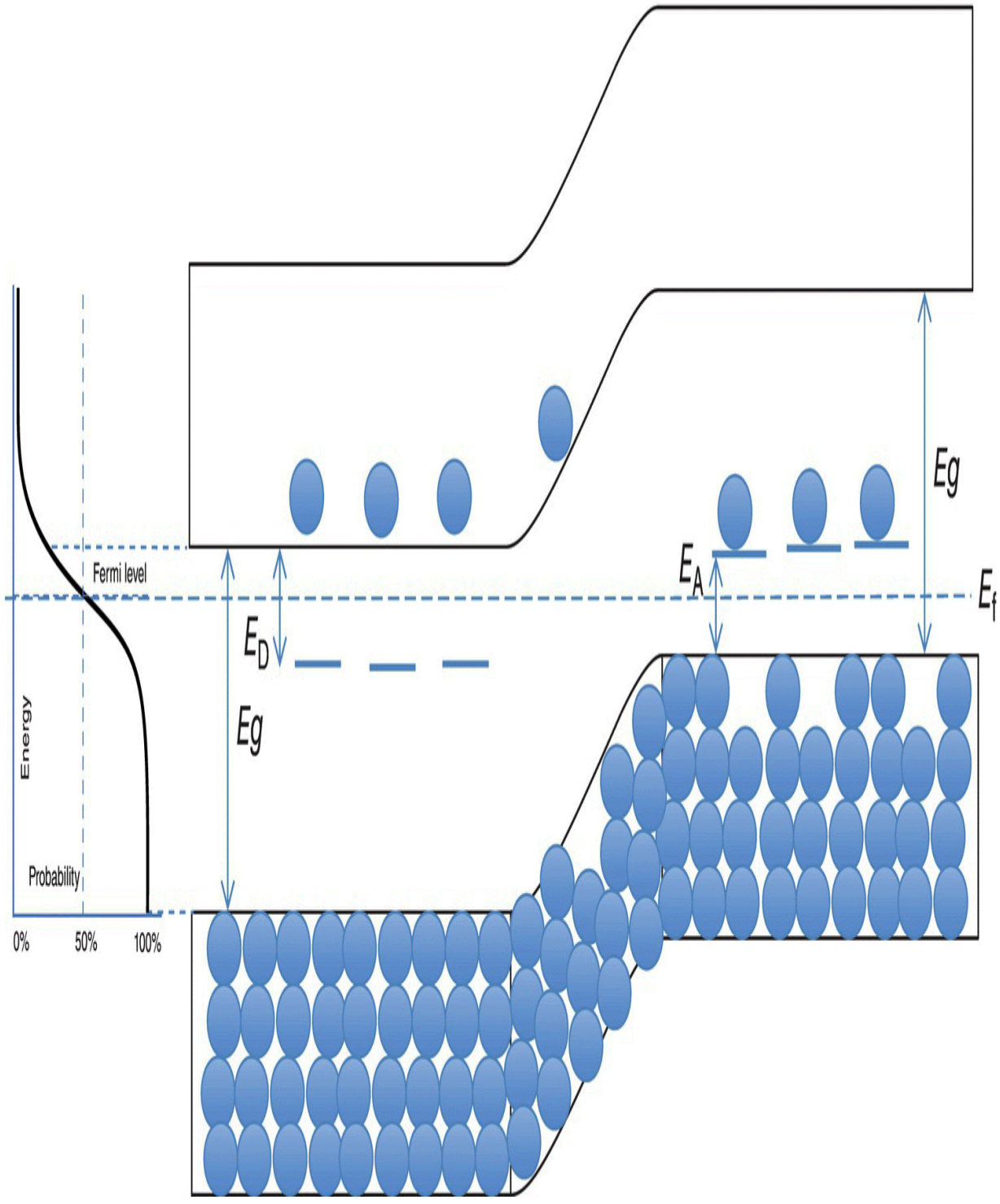

Fermi level
Energy
Probability
0%
50%
100%
E_D
E_g
E_A
E_g
E_f

Figure 5.17 The same pn-junction as in Figure 5.16 but now at 300 K it has electrons in the conduction band of the n-type semiconductor and has trapped electrons in the acceptor levels of the p-type semiconductor, leaving free holes in the valence band. The 300 K F-D function is on the left.

Appendix 5.2 Diffusion and Drift Currents

I mentioned above while explaining the transfer of electrons from the n-type to the p-type semiconductors (and holes the other way around) that the tendency of electrons to move to one side is due to their difference in density and it will stop when an electrical force in the opposite direction cancels it.

The current due to the difference in electron and hole densities, that is, the diffusion current, is

$$\text{diffusion current} = \text{constant} \times \text{charge} \times \text{change in number of electrons } as \text{ a function of } x$$

or

$$i_{nD} = D_n q \Delta n \tag{5.1}$$

where i_{nd} is the diffusion current for the n-type material, q is the charge of the electron, D_n is the diffusion constant for electrons, and Δn is the change in the number of electrons as a function of position. (I use the symbol Δ to indicate the concept of change. Some of you who know calculus will recognize this as the derivative of the number of electrons as a function of position or dn/dx.) The faster the number of electrons change from one position to the next, the larger the current. I can write a similar equation for holes. The diffusing constant for electrons in silicon at room temperature is 93 $cm^2\ s^{-1}$ and for holes is 31 $cm^2\ s^{-1}$. As I mentioned before, the holes are always slower that the electrons.

The second current is the drift current due to the electric fields. The electrons are negatively charged particles so, if I apply a voltage, the

electrons move to the positive terminal.

Mathematically I can write

$$i_{nE} = -q\mu_n n\Delta E \tag{5.2}$$

where i_{nE} is the drift current of electrons, μ_n is the mobility of the electrons, n is the number of electrons, and ΔE is the change in the electric field, the field being the voltage divided by the distance or $E = V/d$. The minus sign, again by convention, is because the electron current is negative. I mentioned the mobility in Section 2.6. For electrons the mobility is 1200 cm^2/volt-s and 250 cm^2/volt-s for holes.

The total electron and hole currents in the semiconductor are:

$$i_n = i_{nD} + i_{nE} = D_n q\Delta n + q\mu_n n\Delta E \tag{5.3}$$

$$i_p = i_{pD} + i_{pE} = D_p q\Delta p - q\mu_p p\Delta E \tag{5.4}$$

In equilibrium, that is no external voltages, the total currents are zero, thus the diffusion current has to be equal to the drift current.

Appendix 5.3 The Thickness of the Transition Region

The thickness of the transition region is the sum of the transition thickness at the n-type semiconductor and the thickness of the p-type semiconductor. Without a proof, let me state that the thickness of the transition regions is given by:

$$x_n = C\sqrt{\frac{N_A}{N_D\left(N_D + N_A\right)}} \tag{5.5}$$

$$x_p = C\sqrt{\frac{N_D}{N_A\left(N_A + N_D\right)}} \tag{5.6}$$

where x_n and x_p are the thicknesses of the transition region in the n and p sides, respectively, and C is a constant for a given device based on the permittivity and the internal voltage of the pn-junction. First assume that $N_A = N_D$, then

$$x_n = x_p = C\sqrt{\frac{1}{2N_D}} \tag{5.7}$$

You can then see that as the doping concentration increases, the transition region gets smaller. If N_A, for example, is much larger that N_D, then,

$$x_n \approx C\sqrt{\frac{1}{N_D}} \quad \text{and} \quad x_p \approx C\sqrt{\frac{N_D}{N_A^2}} \tag{5.8}$$

Let's put in some numbers to see what that means. For a typical donor concentration such as I have been using, $C = 3.3$ m. Let's say $N_A = 10^{18}$ and $N_D = 10^{16}$, then

$$x_n = C\sqrt{\frac{1}{10^{16}}} = 3.3\times 10^{-8}\,\text{m} = 330\,\text{nm} \tag{5.9}$$

and

$$x_p = C\sqrt{\frac{10^{16}}{(10^{18})^2}} = 3.3\times 10^{-10}\,\text{m} = 3.3\text{nm} \tag{5.10}$$

The transition region width is 100 times larger in the n-type region than in the p-type semiconductors. This actually makes a lot of sense. Look at Figure 5.18. If the density of holes in the p-type semiconductor is half that of the density of electrons in the n-type semiconductor we will need to go twice as far into the into the p-type region to get all the charges we need to create the pn-junction, so x_p will be twice as long as x_n. That is why we want the concentrations to be very high in tunnel diodes, so that the transition region is as narrow as possible.

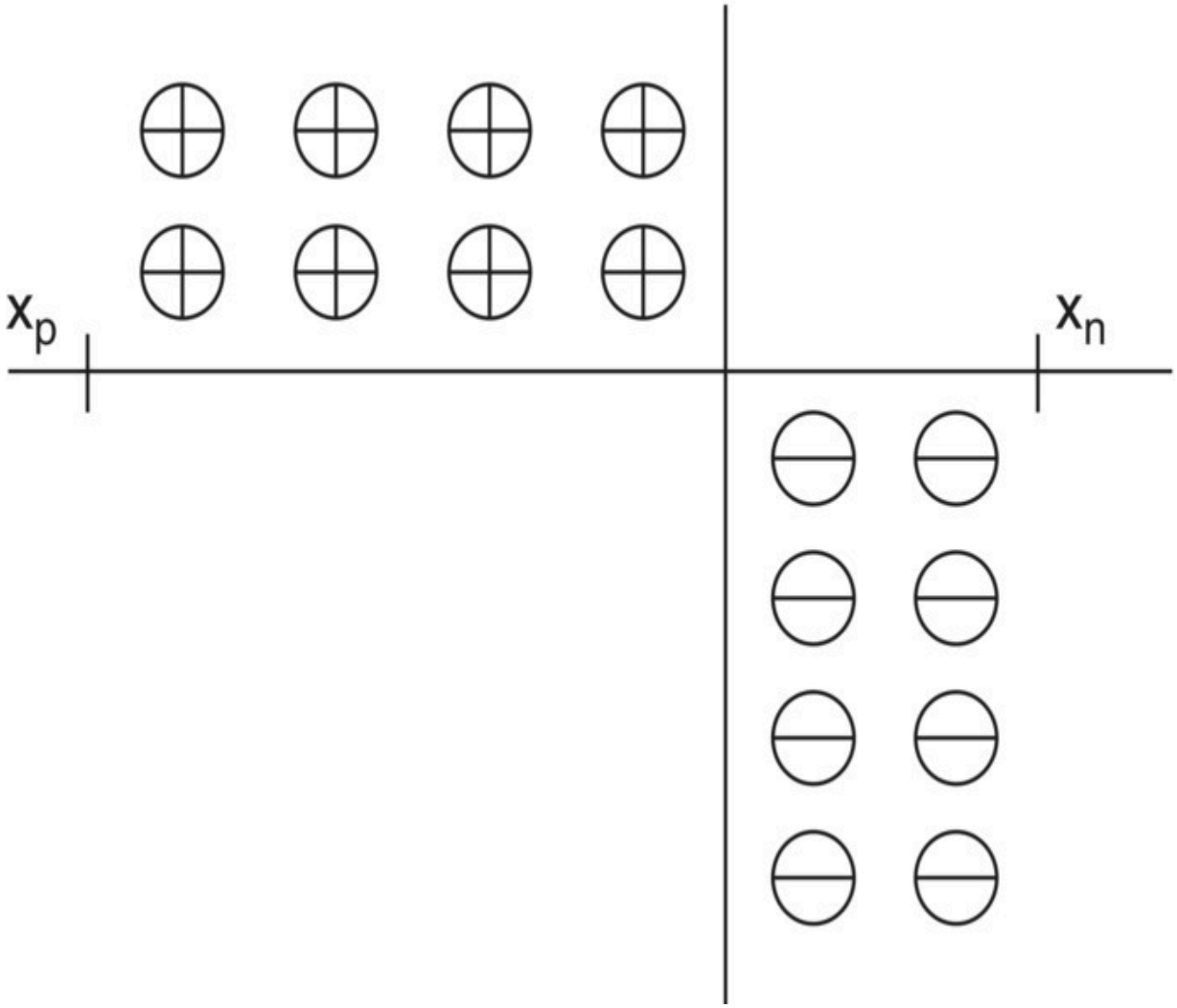


Figure 5.18 If the p-type semiconductor has half the concentration of impurities as the n-type, we will have to go twice as far into the p-side to get all the charges we need in the n-side to create the depletion region.

Appendix 5.4 Work Function and the Schottky Diode

To understand how the Schottky diode works, I need to explain another property of materials, their work function. If I give enough energy to a material, not only I would free electrons so they can move freely inside the material, I can actually knock an electron out of the material altogether. This is what a cathode-ray tube does.

The work function is the amount of energy an electron in a material needs to get to the surface of the material and out of the solid so it can be swept away by an electric field, if there is any. This is the way old TV sets (remember the big boxes) used to work.

Look at part A in Figure 5.19. In a metal there is no separation between the valence band and the conduction band. At 0 K, all the electrons take the lowest possible energy the same way as the water in a pot delineates a perfect surface as long as we do not shake it or boil it, that is, it has no energy. The quiet surface is the Fermi level in the metal. At 300 K, which is what I show in part A in Figure 5.19, there is energy and some of the electrons move above the Fermi level leaving behind equal empty spaces. If the energy is high enough, some electrons will jump to the surface and escape the metal all together. This energy is what we call the *work function* of the metal, W_M. Aluminum, for example, has a work function of 4.2 eV. (This value is not really a constant. It changes with composition, crystal orientation, and surface properties.)

Now look at part B in Figure 5.19. We are already familiar with an n-type semiconductor at 300 K (Appendix 3.1, Figure 3.17). The Fermi level for a n-type semiconductor sits between the conduction band and the donor levels because at 0 K all the levels below the Fermi level are occupied and all the levels above are empty. At 300 K, all the donor electrons have sufficient energy to move to the conduction band. This is the situation I show in Figure 5.19, part B. It takes additional energy for the electrons in the conduction band to escape the material altogether. I call this escape energy W_S and for an n-type silicon it is 4.8 eV.

What happens when we grow a metal on top of the n-type semiconductor? Notice first in Figure 5.19 parts A and B that the electrons in the conduction band of the semiconductor have a higher energy than the electrons in the metal compared to the common reference, the vacuum level. So there is a desire for the electrons in the semiconductor to jump to the metal (I show this as a dashed arrow between B and A). When I bring the two materials into

intimate contact, the two different materials form a single system, and therefore the Fermi levels have to be the same (Figure 5.19, part C). The electrons in the conduction band in the semiconductor have a higher energy than the electrons in the metal and therefore there is a movement of electrons from the semiconductor to the metal. As in the case of the regular pn-junction, the semiconductor loses electrons near the junction and becomes positively charged and the electric field generated eventually stops any further electrons from moving to the metal. This disturbs the bands at the interphase but the properties of the semiconductor away from the interphase remain the same. There are electrons moving back and forth across the barrier in both directions but there is no net current. As I mentioned in Section 5.1 and in more detail in Appendix 5.3, the transition region thickness is inversely proportional to the number of free electrons. The same is true here: the transition region is all located on the semiconductor side and not on the metal side, which has thousands more electrons per unit area than the semiconductor.

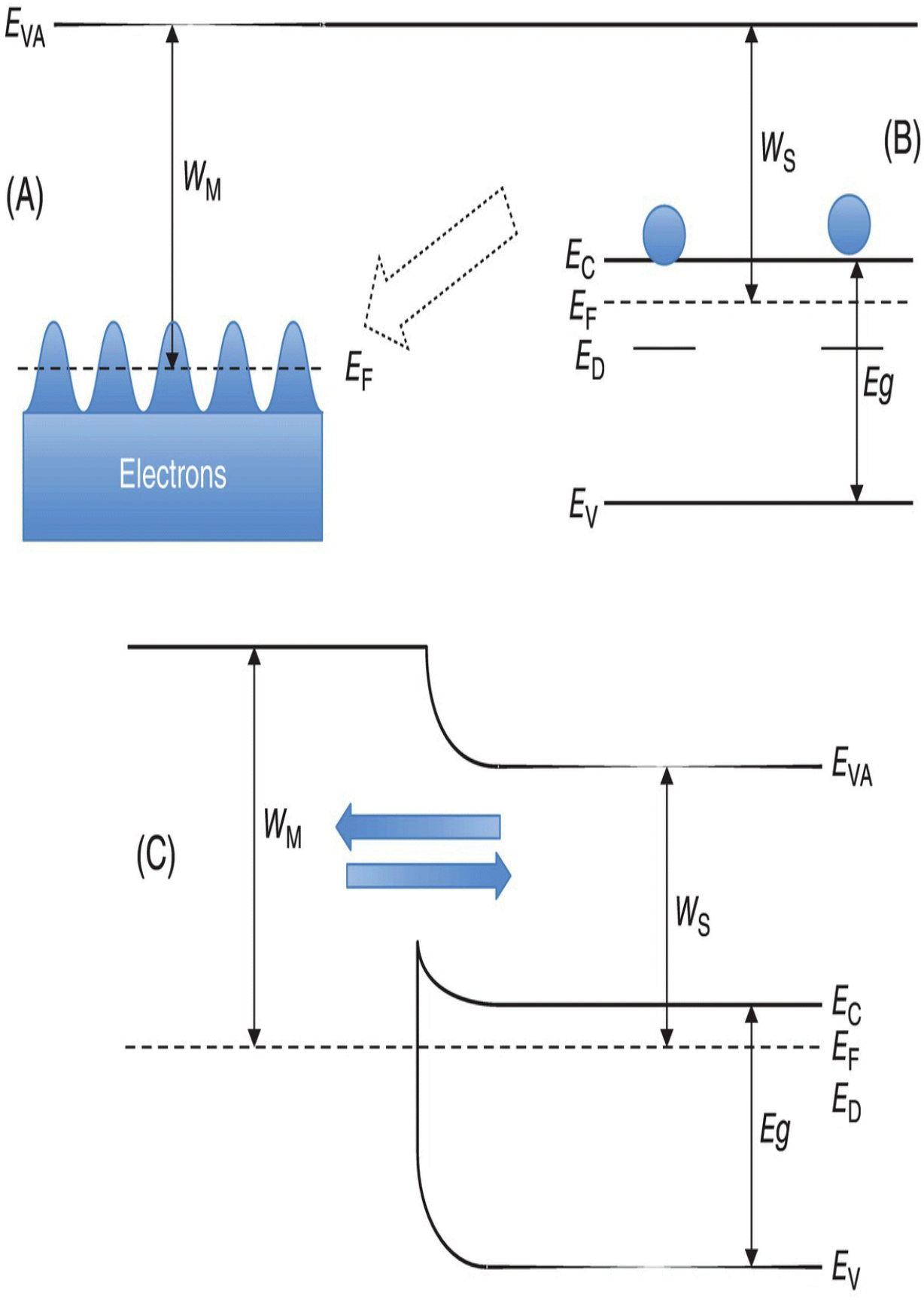

E_{VA}
(A)
W_M
E_F
Electrons
W_S
(B)
E_C
E_F
E_D
Eg
E_V
(C)
W_M
E_{VA}
W_S
E_C
E_F
E_D
Eg
E_V

Figure 5.19 The vacuum level E_{VA} is the same for all materials. The Fermi levels at room temperature in a metal (A) and in an n-type semiconductor (B) have to align when they are in intimate contact (C), generating an energy barrier.

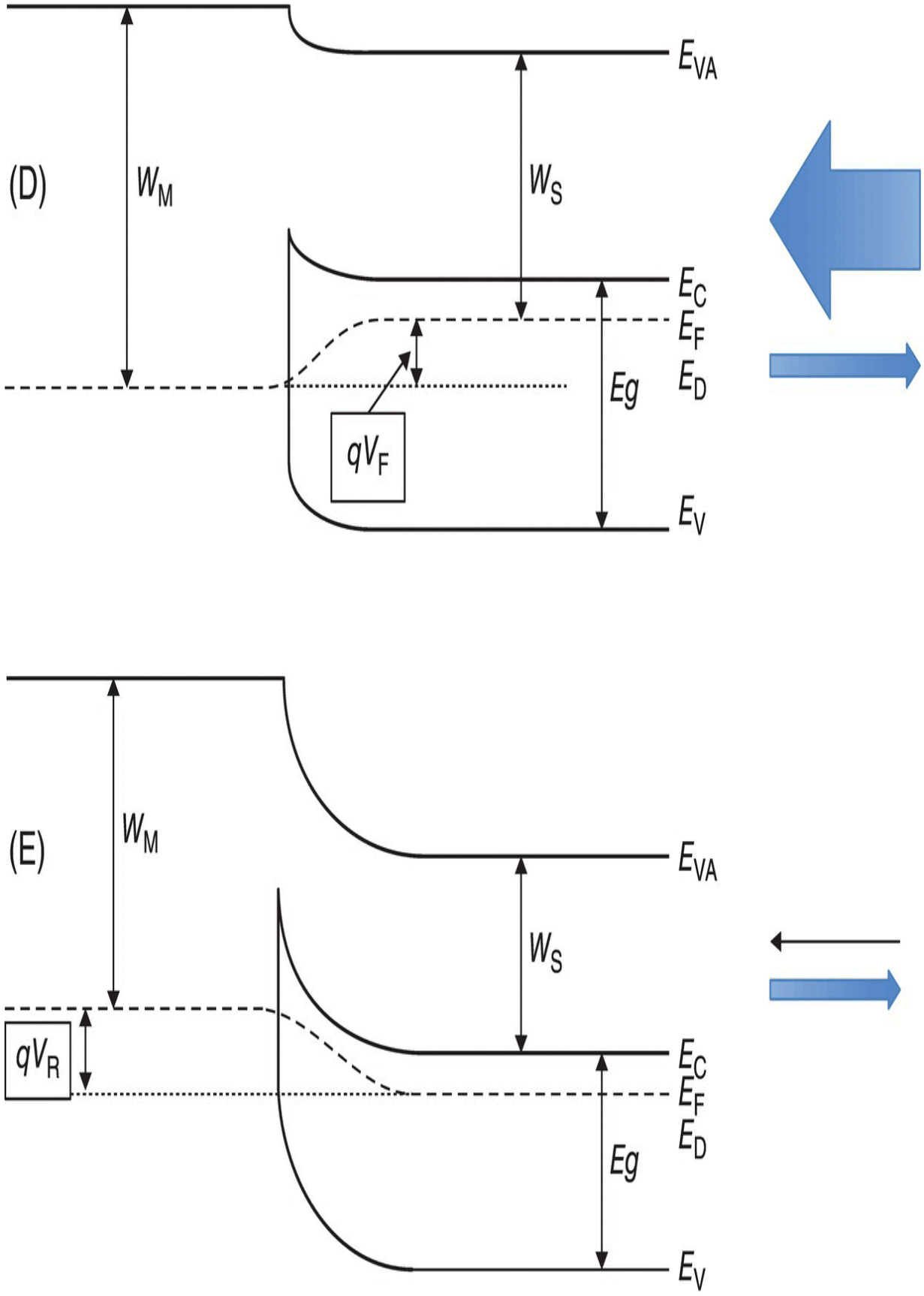

(D)
W_M
W_S
E_{VA}
E_C
E_F
E_D
Eg
qV_F
E_V
(E)
W_M
W_S
E_{VA}
qV_R
E_C
E_F
E_D
Eg
E_V

Figure 5.20 The Shockey diode under the forward bias condition (D) the barrier decreases, and electrons flow from the semiconductor to the metal and in reversed bias condition (E) the barrier increases, stopping the flow. The current from metal to semiconductor remains the same.

When we apply a voltage to the Schottky diode, we have the same behavior as in the regular pn-junction diode, as I show in Figure 5.9.

If you compare parts C, D, and E in Figures 5.19 and 5.20, you will notice that the barrier as seen from the metal side has not changed but the barrier as seen from the semiconductor side has decreased in part D and increased in part E. In the forward biased case, the barrier from the semiconductor to the metal has decreased by an amount qV_F, F for forward bias, so the current going from the semiconductor to the metal is now larger than the one going from the metal to the semiconductor as I show in part D of Figure 5.20.

In Figure 5.20 part E, I show the reverse biased case. Now the electrons from the semiconductor see a very high barrier they have to overcome to go to the metal and therefore the only electron movement is the small residual current from the metal to the semiconductor.

The interesting thing, as you look at Figure 5.20, is that the current from the metal to the semiconductor is small and does not change, but is the same under forward and reverse bias. The diode performance is only due to the barrier the semiconductor sees. This, by the way, is one of the problems with the Schottky diode: even though it is faster and uses less power than the semiconductor diode, the reversed bias current is larger.

6
Other Electrical Components

OBJECTIVES OF THIS CHAPTER

After explaining how the diode operates, I was eager to start discussing some of its applications. I realized, though, that there are other electrical passive components that are necessary to design useful electronic circuits and we need to understand them. Therefore, after a digression on voltage and current, both direct and alternating, I discuss three electrical/electronical components: the resistor, the capacitor, and the inductor. The first two are indispensable to creating working electronic circuits and systems. Inductors are not as commonly used, partly because it is awkward to fabricate them in a miniaturized setting. Some of you may already be familiar with these elements and may want to scan or skip this chapter.

6.1 Voltage and Current

First, just a comment about voltage and current, which sometimes are confused. Let me explain the difference using the analog of fluids. As a matter of fact, there is a field call fluidics that mimics electronic circuits, but it uses liquids, fluids, instead of electrons.

To have motion of fluids you need, obviously, the fluid. You also need a path for the fluid to move through and a force, a pump or just gravity, to push the fluid through the chosen path. The same is true in an electrical system. The battery or the generator is the equivalent of the pump. The batteries sitting in your drawer have an electrical potential which we call voltage and the units are measured

in volts, the same way that the water can be stored up in a water tank with sufficient potential energy to come down as soon as we provide a path. In the case of a battery, the voltage is typically 1.5 V. The energy provided by the chemical composition is inside the battery no matter if the battery is or is not connected. If I connect the battery to a light bulb, as in a flashlight, current flows and the light is on until the energy stored in the battery is completely spent (after many uses you need to replace the batteries). The same thing happens with a water tank. The water will stay up there until we provide a path and the water continues flowing until the tank is empty.

The current is the motion of electrons the same way as the flow is the motion of water. How much water flows depends on how powerful the pump is and how easy is the path that the water has to follow. Similarly, the current, the flow of electrons, depends on how strong the battery or the generator is and how easy is the path through which the electrons have to flow.

The flow of water is measured by gallons per second (or liters, or quarts per second). The current, *I*, has units of amperes, *A*, and is the electrical charge per second, measured in coulombs, *Q*, per second where the coulomb, like the gallon, is a measure of the quantity of electrons (one coulomb is the charge of 6.242×10^{18} electrons), and you already know that the power of the battery or generator is measured in volts.

6.2 Resistance

The resistance, *R*, is the measurement of the difficulty that a specific material presents to the flow of current.

There is an equivalent to the resistance in fluidics. Take a look a Figure 6.1. The simple fluidic circuit on the left of Figure 6.1 consists of a pump connected by a pipe to a sandbox. If the sandbox is full of large rocks or very coarse sand, the pump does not have much difficulty pushing the water though it and the water flowing in the

circuit is high. But if the box is filled with very fine sand, the resistance to the water flowing through it is higher and the flow of water throughout the path is smaller. For a given pump power, the flow of the water decreases as the sand gets finer and finer (or the resistance to the water increases). If I want the same flow of water through the circuit when I have a box full of fine sand, I have to increase the power of the pump. How much water flows is therefore proportional to how high the power of the pump is, and inversely proportional to how resistive the sandbox is to the flow of water.

The resistance of the sandbox to the flow of water depends on the following:

The properties of the sand inside the sandbox. The finer the sand, the higher the resistance, and thus less fluid flows.

The length, *L*, of the box. The longer the box, the more resistance to the flow of water.

The area: the wider the box, the more ways the water finds to flow through it and the resistance is less.

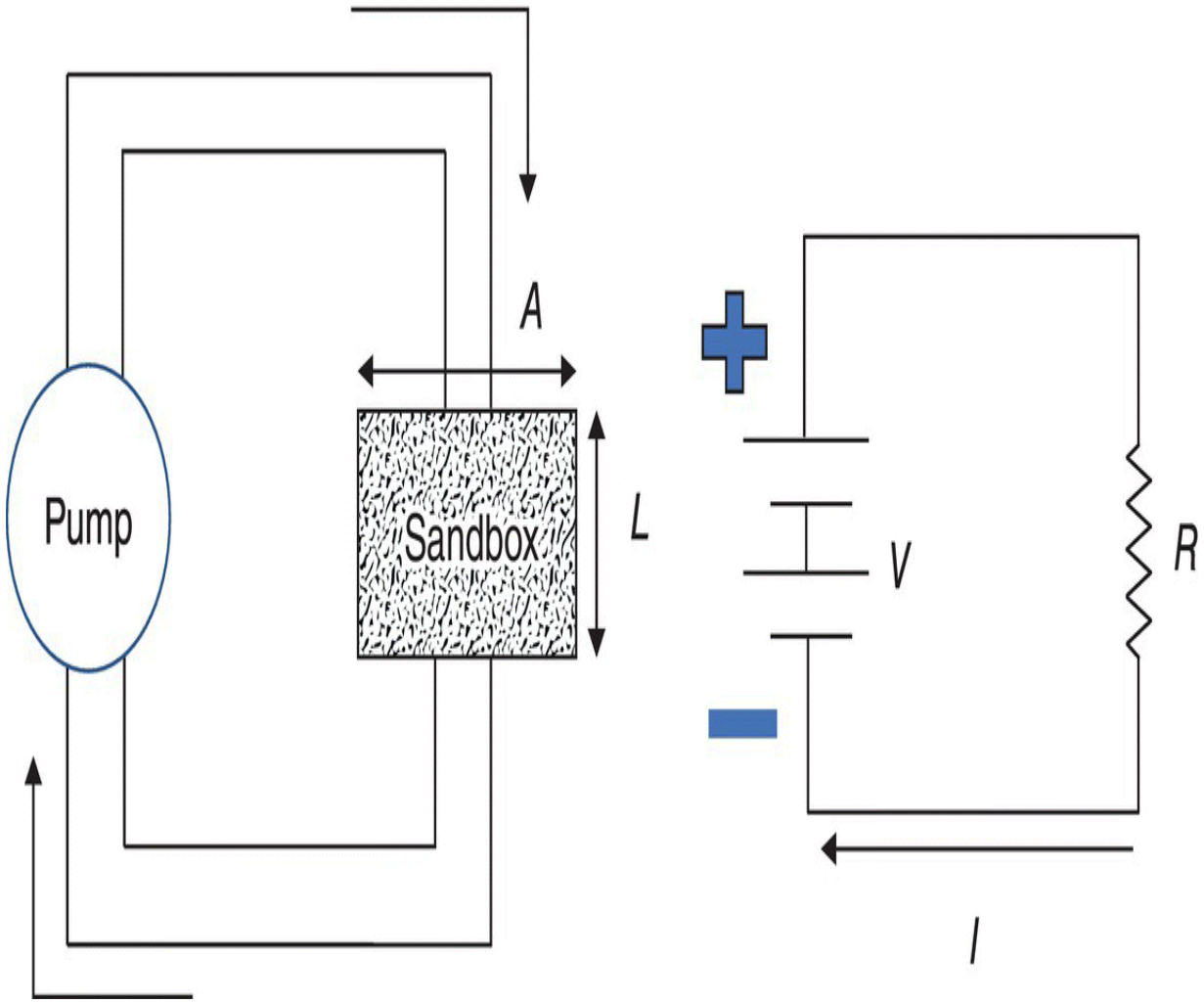


Figure 6.1 A fluidic analogue of an electrical circuit with resistance to the flow of water (left) or electrical current (right).

I also would point out that the power the pump needs is a function of the amount of water flowing and the speed of the water. The more these two quantities increase, the higher the power of the pump needs to be.

These fluid concepts are almost identical to those for the electrical circuit that I show schematically on the right of Figure 6.1. The current, I, is proportional to the power of the battery or generator, and inversely proportional to the resistance of the element attached to the source.

We measure the resistance in ohms and the standard symbol for ohms is the capital Greek letter omega, Ω.

There is a very simple relationship between the voltage, the current, and the resistance, known as Ohms law:

$$\text{voltage} = \text{current} \times \text{resistance}$$

$$V = IR \tag{6.1}$$

The power dissipated in the resistance is, similarly to the fluid case,

$$\text{power} = \text{voltage} \times \text{current}$$

$$P = VI \tag{6.2}$$

By combining the two relationships, Eqs. (6.1) and (6.2), I get

$$P = VI = I^2 R \tag{6.3}$$

Here is a very simple problem (Oops, I said in the introduction I would have no problems in this book. Please call the problem an "example"): What is the resistance of a 150 W lamp connected to the electrical 120 V home line?

The current is

$$I = \frac{P}{V} = \frac{150\,\text{W}}{120\,\text{V}} = 1.25\,\text{A} \tag{6.4}$$

and therefore the resistance is

$$R = \frac{V}{I} = \frac{120\,\text{V}}{1.25\,\text{A}} = 96\,\Omega \tag{6.5}$$

The resistance of a material is proportional to the length and inversely proportional to the area of the material, the same as the resistance of the sandbox to the water flow. This makes a lot of

sense. Given a material with certain resistance, if I double its length the resistance should increase by a factor of two and if I increase its thickness also by a factor of two, the resistance should decrease by a factor of two. The thinner the material, the higher the resistance. A very simple relation describes this behavior

$$R = \rho \frac{L}{A} \tag{6.6}$$

where L is the length and A is the area of the resistor. This new Greek letter rho, ρ, is called the resistivity of the material and has the units of ohm meters, or Ω-m. This number is practical because it is a constant for each material and does not depend on its dimensions. The range of resistivities changes drastically for different materials, from silver with a resistivity of 1.59×10^{-8} Ω-m, to diamond (carbon) with a resistivity of 1×10^{12} Ω-m, a range of 20 zeros. The semiconductors are very much in the middle range with intrinsic resistivities around 1 Ω-m. The resistivity, though, changes with temperature. The numbers above are at room temperature (20 °C) and the resistivity of a semiconductor also changes drastically, both with temperature (Figure 2.9) and with the number of added impurities (Figure 3.12).

Resistances can be combined in series and in parallel (Figure 6.2).

The two resistances on the left of the Figure 6.2 are in series. The same current is flowing through both resistors. The total combined resistance is just the sum of both resistors,

$$R_T = R_1 + R_2 \tag{6.7}$$

The current, I, is the voltage divided by the total resistance

$$I = \frac{V}{R_1 + R_2} \tag{6.8}$$

Now the voltage V_A is just the current times the resistance R_2:

$$V_A = IR_2 \tag{6.9}$$

Combining Eqs. (6.9) and (6.10) we end up with

$$V_A = V\frac{R_2}{R_1 + R_2} \tag{6.10}$$

This is the concept of a voltage divider. I can get any voltage V_A I want, as long as it is smaller than V, by just selecting different values of resistances. As a matter of fact, any one of the resistances could be a variable resistor, a potentiometer, and V_A could have any value between 0 and V/R_1.

The drawing on the right of Figure 6.2 is a current divider. The voltage V is the same for both resistors and the total current is the sum of the two currents, so

$$I_1 = VR_1 \quad \text{and} \quad I_2 = VR_2 \quad \text{and} \quad I = I_1 + I_2 \tag{6.11}$$

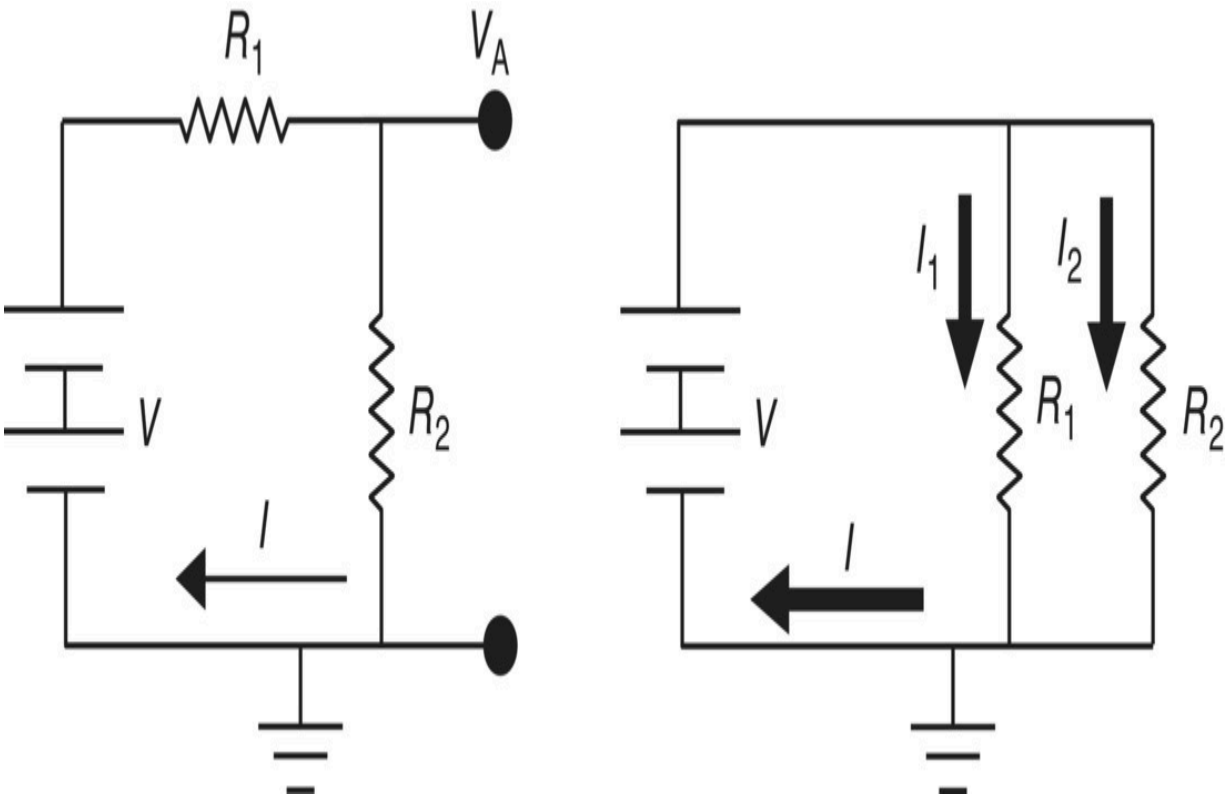


Figure 6.2 Resistors in series (left) divides the voltage and in parallel (right) divides the current.

Therefore, the total equivalent resistance, R_T, that the battery, or the voltage source, sees is

$$I = \frac{V}{R_1} + \frac{V}{R_2} = V\left(\frac{1}{R_1} + \frac{1}{R_2}\right) = V\frac{R_1 + R_2}{R_1 R_2} = \frac{V}{R_T} \tag{6.12}$$

Therefore, the voltage source sees an equivalent total resistance of

$$R_T = \frac{R_1 R_2}{R_1 + R_2} \tag{6.13}$$

If either R_1 or R_2 are zero, R_T is also zero since the battery is shorted in at least one of the paths. By selecting different resistance values, I can divide the total current and decide which of the

currents dominates. If both resistances are the same, each of the currents is one half of the total current, which is what we expect just by looking at Figure 6.2.

I will use these concepts later on in the book.

6.3 The Capacitor

The next most common component in electronics is the capacitor. Continuing with the fluidics analogy, I can compare the capacitor to a box with a flexible but no porous membrane (Figure 6.3).

When the pump is off, Figure 6.3A, the membrane is horizontal. The water applies the same pressure on top as at the bottom of the membrane. But when the pump is turned on, the water accumulates at the top of the membrane distorting its shape, Figure 6.3B. Think about it. The water flows until (i) the pump is turned off or (ii) the elastic force of the membrane counterbalances the force of the pump. When I turn the pump off, the water does not flow anywhere so the membrane remains distorted, as in Figure 6.3B.

Let me show you how this membrane works on a fluidic circuit. Look at Figure 6.4. Let's go one step at the time. Take a look first at the top drawing of Figure 6.4. As I turn the pump on, instantaneously there is a constant flow of water through the sandbox limited by the strength of the pump and the resistivity of the sandbox. At the start, there is also a rush of water to the membrane until the membrane cannot stretch anymore. At this point, the flow through the capacitor goes down to zero. No more water flows to the membrane. The flow through the sandbox continues uninterrupted. (The arrow at the top of the figure indicates the flow of the water. I show a triangular shape going into the capacitor to indicate the initial water flow is high and it dies down after the capacitor is full. After the membrane box is fully stretched, the current continues flowing only through the sand box.)

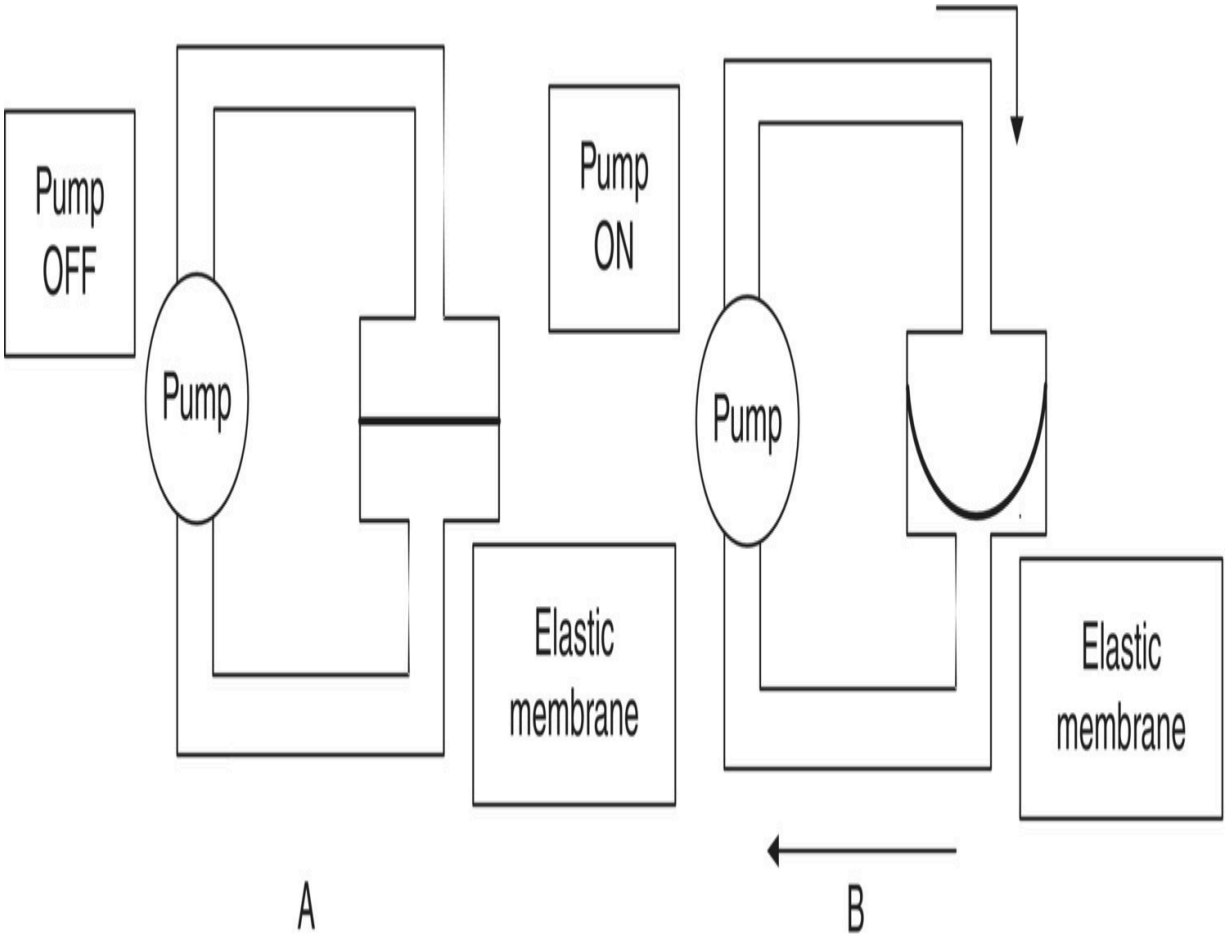


Figure 6.3 A flexible membrane stores water. Water flows almost instantaneously until the membrane cannot stretch any longer.

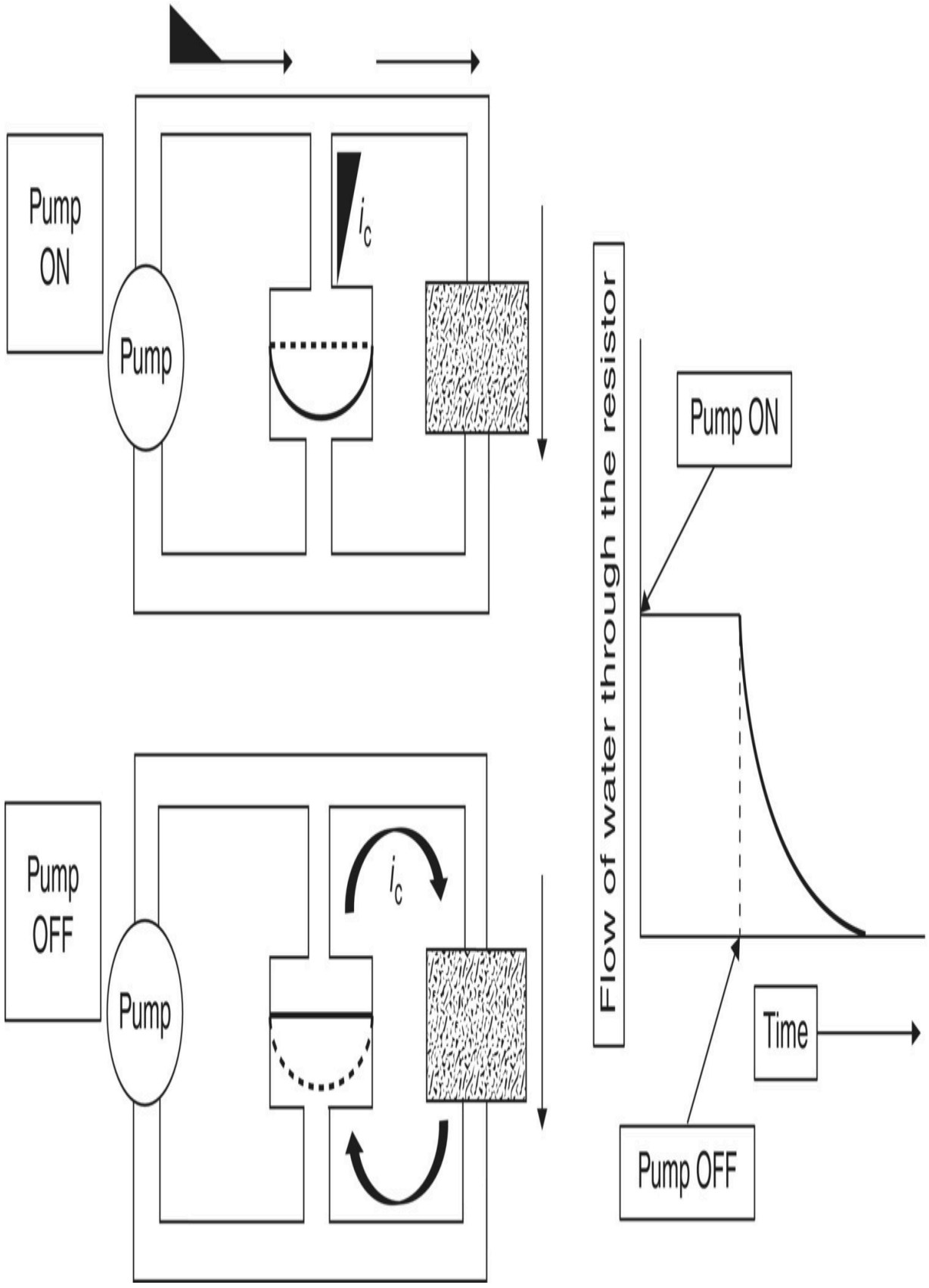

Pump
ON
Pump
i_c
Pump
OFF
Pump
i_c
Flow of water through the resistor
Pump ON
Time
Pump OFF

Figure 6.4 When I turn the pump on, there is current through the sand box and initially into the capacitor until the membrane cannot store any more water (upper figure). When the pump is turned off, the capacitor sends the accumulated water through the sand box until there is no water left in the membrane box (lower figure). The plot on the left shows the water flowing through the sandbox.

Next, let me turn the pump off. No more water is flowing from the pump. The membrane pushes the water back and sends its stored water through the resistor until all the water stored in the membrane is gone. At the right of Figure 6.4 I show the plot of the flow of water through the sandbox only. As I turn the pump on, the water flows uniformly through the sandbox, but when I turn the pump off, instead of the flow going instantaneously to zero, the water stored in the membrane continues to flow through the sandbox until all the water stored at the membrane is gone.

This is exactly what an electrical capacitor does.

The electrical capacitor is basically a sandwich composed of an insulating material between two conductive plates (Figure 6.5).

If I turn on the switch, exactly as in the fluidic case, the current instantaneous flows through the resistor and the capacitor charges to the voltage V. The current, V/R, flows through the resistor as long as the switch is on, but when I turn the switch off, the charges accumulated in the capacitor discharge through the resistor and the current continues flowing until no more charges remain in the capacitor. I show the current through the resistor at the right of Figure 6.5 and its shape is identical to the fluidic case in Figure 6.4.

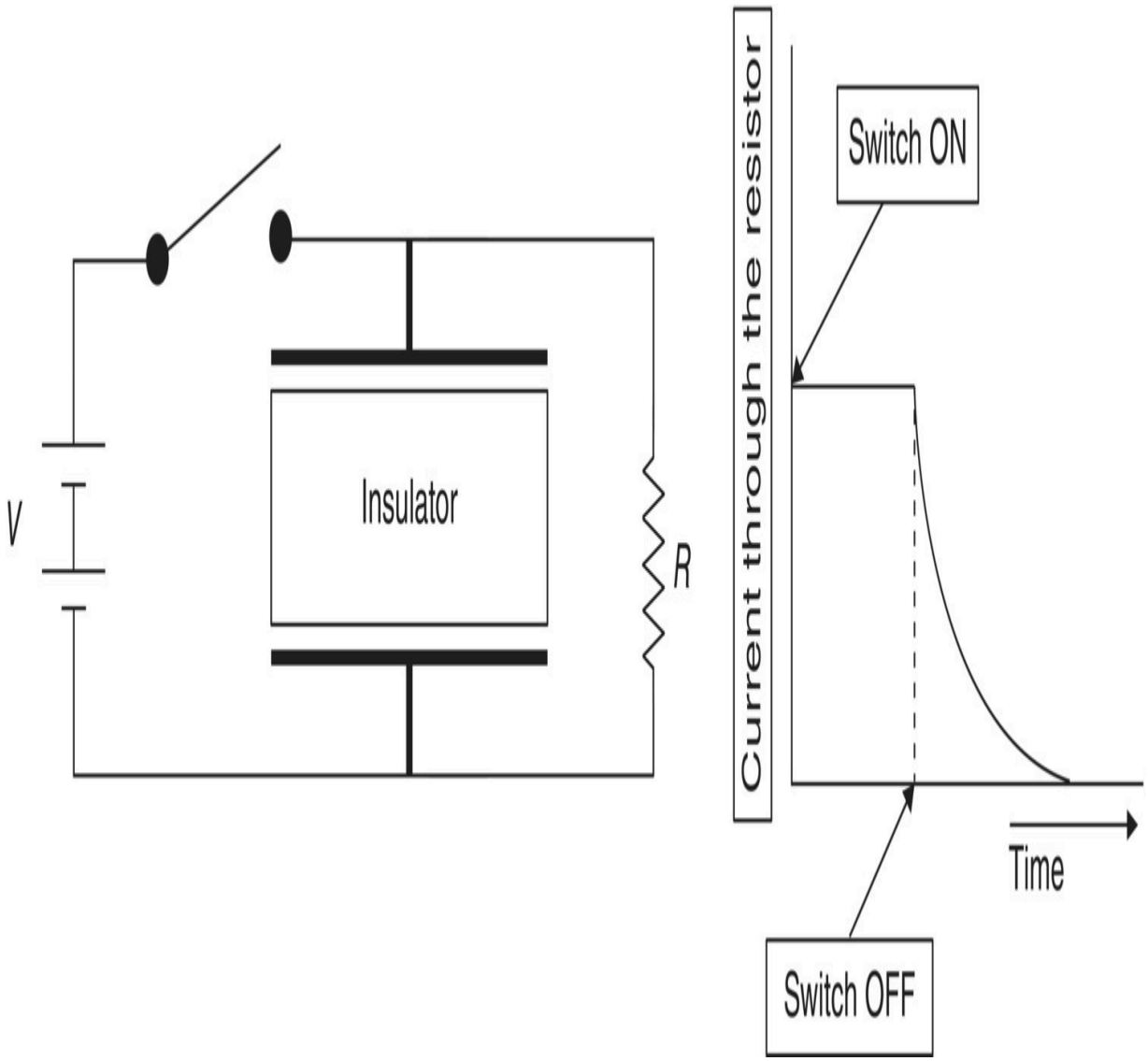


Figure 6.5 A capacitor consists of two parallel plates separated by an insulator. When a voltage is applied, the plates hold charge until its voltage equals the source voltage. When we open the switch, the capacitor discharges though the resistor, generating a decaying current.

Have you ever wondered why, when you have a problem with an electronic appliance, the technician asks you to unplug the appliance, wait a few seconds, and plug it back in? This mysterious solution that resolves up to 50% of problems is because we want to give the system time to discharge all the capacitors in the circuit.

How many charges the capacitor can hold depends upon its area, the type of insulating material, the thickness of the insulating material, and the applied voltage. We can express this relation by:

$$Q = CV \tag{6.14}$$

where Q is the number of charges measured in coulombs and C is the value capacitor. As in the case with the resistance, we can also talk about a *capacitance* which, similar to the resistivity, is a property of the insulating material. This capacitance allows us to calculate the value of the capacitor if we know its dimensions and the insulating material we use. The relationship is

$$C = \frac{\varepsilon_o \varepsilon_r A}{d} \tag{6.15}$$

where A is the area of the conductive plates, d is the separation between the plates, ε_0 is the permittivity of free space ($\varepsilon_0 = 8.85 \times 10^{-12}$ m^{-3} kg^{-1} s^4 A^2), and ε_r is the relative permittivity, that is, the permittivity of the specific material compared to the permittivity of free space. The units of the capacitor are Farads and the units of the permittivity are Farads/meter. The relative permittivity of SiO_2, the most prevalent insulator used in integrated circuit fabrication, is $\varepsilon_{rSiO2} = 3.9$.

The relationship between voltage and current in the capacitor is given by

$$i = C\Delta V \tag{6.16}$$

I use the symbol Δ here because there is current in the circuit *only* when the voltage changes. Think about it. Even though the circuit is interrupted by an insulating material between the two plates, at the moment I connect the battery to the capacitor there is a rush of electrons to one plate, making it negative, and a reduction of electrons in the opposite plate, making it positive. For an external observer, it looks like there is current through the capacitor even

though no charges move between the two plates. But as soon as the potential at the plates is equal to the voltage at the battery, the current stops because the voltage does not change. So there is current only when the voltage across the capacitor changes.

6.4 The Inductor

An inductor is basically a coil made of a conductive material (Figure 6.6, left). A current flowing through the coil generates a magnetic field which stores the electrical energy. This magnetic field is proportional to the number of turns, the cross-section of the coil, and the magnetic properties of the material inside the coil.

The inductor is equivalent to the water wheel attached to a fly wheel. As the water start flowing, the water wheel starts rotating, first very slowly and then faster and faster until its speed is the same as the speed of the water. How much time it takes for the water wheel to reach its steady-state speed depends upon the inertia of the wheel, that is, if I attach a very heavy fly wheel to the axis of the water wheel, it will take a long time for the wheel to start moving fast. When the wheel reaches equilibrium and it is turning as fast as the water flow, the wheel poses no resistance whatsoever to the flow of water, very similar to the inductor in an electric circuit (assuming again that there is no friction).

Now let's see what happens when we add a resistor like we did with the capacitor. Figure 6.7 shows this case. Note that I have replaced the membrane by the water wheel and the sandbox is now in series with the water wheel.

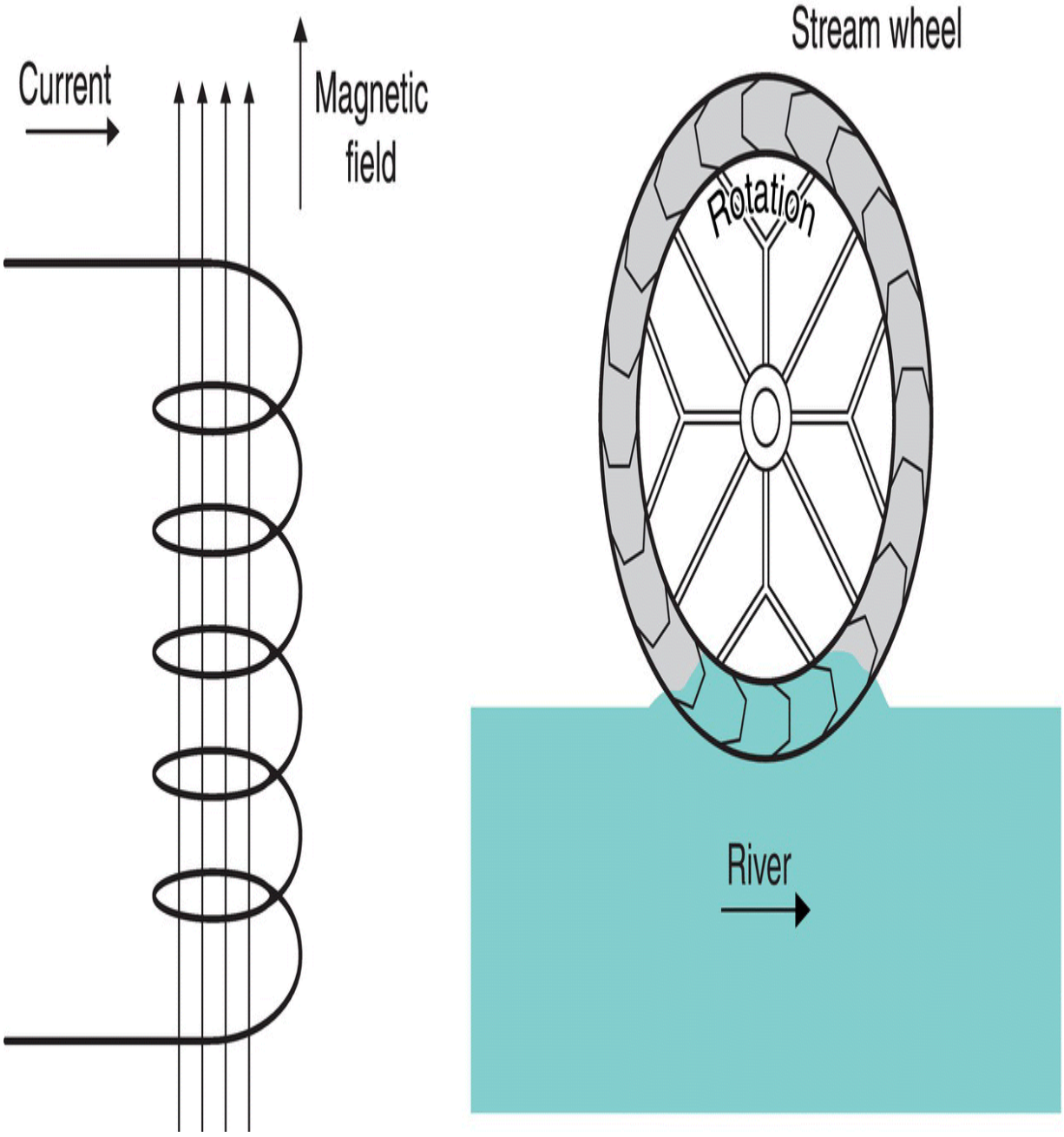


Figure 6.6 An inductor stores electric energy in the form of a magnetic field in a similar way to a water wheel storing kinetic energy.

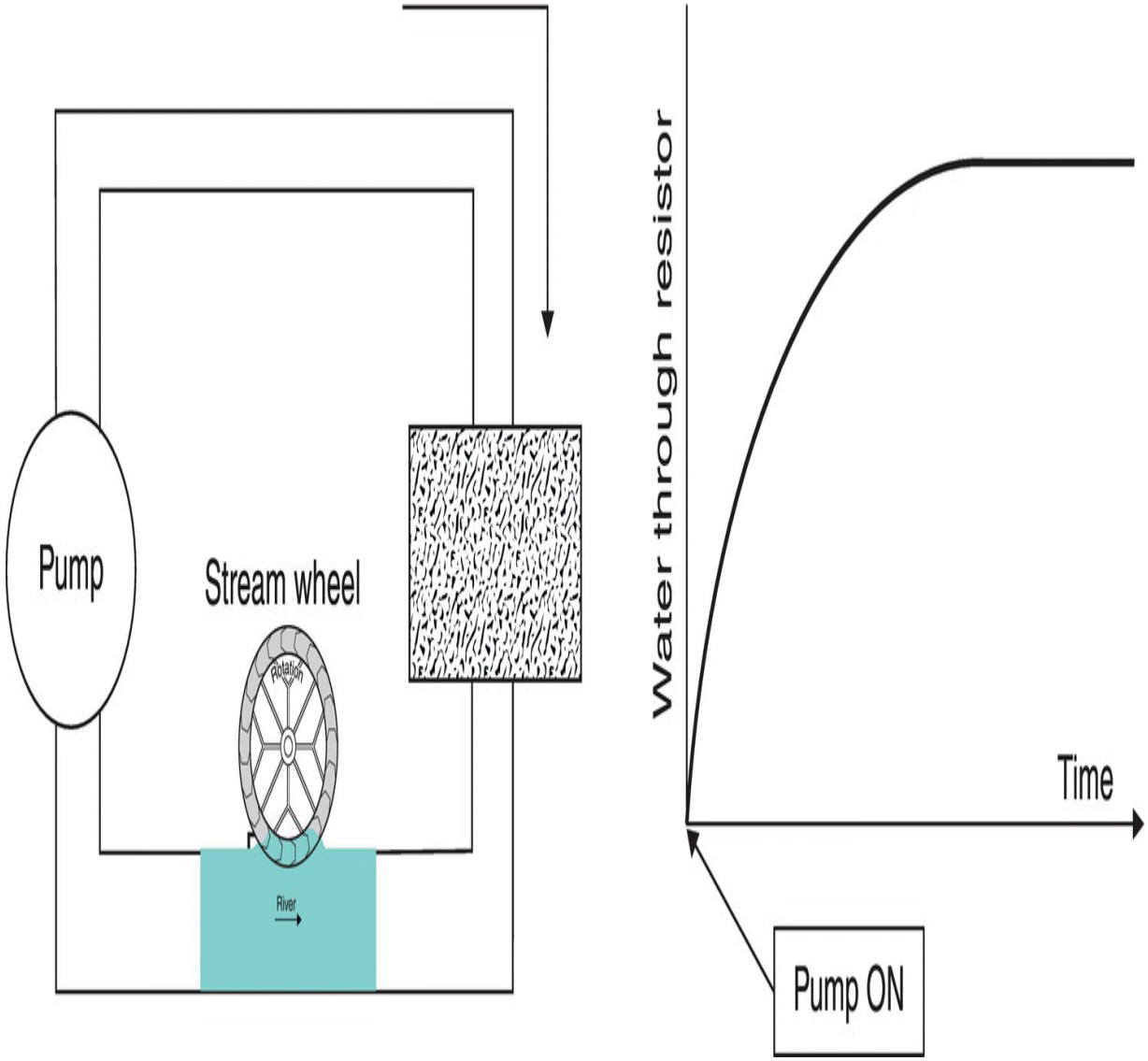


Figure 6.7 When the pump is turned on, the water wheel starts moving, first slowly and then, as the speed increases, the water flow through the box increases to a value limited only by the resistance of the sandbox.

Now the current through the elements is similar but different. When we first turn the pump on the water wheel with a fly wheel resists the flow of water and it slowly starts moving. Therefore, the flow through the sandbox starts at zero and increases as the fly wheel gets more and more rotational energy. When it reaches the steady state, the fly wheel does not present any resistance to the flow of

water and the flow depends only on the power of the pump and the resistance of the sandbox, as we saw in Section 6.2.

The same thing is happening with an inductor connected to a battery and a resistor. At the instant of time I turn the switch on, there is no current in the circuit. A magnetic field starts forming and the current increases until the magnetic field is fully formed, and the current is limited only by the resistance ($I = V/R$). The shape of the current is exactly the same as the one I show at the right of Figure 6.7.

What I failed to tell you, on purpose, is what happens if I turn the water, or the switch, instantaneously off. Both the kinetic energy of the fly wheel and the magnetic field of the inductor have to collapse instantaneously. In the fluid system, the pressure of the fly wheel is going to be immense and probably blow the pipe (imagine the catastrophe if we were to stop a train instantaneously). The same thing with the magnetic field. If we turn the circuit off or unplug the circuit, there is an instantaneous huge current. You have observed the spark in the plug when you pull the cord out an electrical socket.

The value of the inductance is calculated, like we have seen with resistors and capacitors, by the relation:

$$L = \mu_o \mu_r N^2 l A \tag{6.17}$$

where L is the inductance, μ_0 is the magnetic susceptibility or permeability of free space ($\mu_0 = 1.26 \times 10^{-6}$ H m^{-1} or kg m s^{-2} A^{-2}), μ_r is the relative permeability, N is the number of turns of the coil, A is the cross-section area of the coil, and l is the length of the coil. The unit of inductance is the Henry (H = kg m^2 s^{-2} A^{-2}). The relative permeability of iron is between 5000 and 200 000, depending on its purity, quite an increase over air, which is only 1.0006, just a tiny bit larger than vacuum.

The voltage across the inductor is proportional not to the current, but to the change in the current, which I can write as:

$$v = L\Delta i \tag{6.18}$$

So there is a voltage across the inductor only if the current through its coils keep on changing. Note that this relation between voltage and current is exactly the opposite of the capacitor. The current exists only if the voltage across the capacitor changes.

There is also another important difference. The plates of the capacitor are separated by an insulator, so there is not physical motion of electrons from one plate to the other through the insulator. At some point the voltage is so large that the dielectric breaks down. When you buy a capacitor, it not only tells you its value (in Farads) but also the maximum voltage you can apply to it.

The opposite is true for the inductor. The wire that forms the coils in any inductor has practically no resistance, so even a small constant voltage generates extremely large currents (remember $I = V/R$) and will burn either the inductor or the source. Don't ever connect a coil to a battery.

6.5 Sinusoidal Voltage

In the 1880s there was a competition between Thomas Edison (1847–1931), who wanted to develop an electrical system based on direct, constant, voltage, and current, and George Westinghouse (1846–1914), who proposed a system of alternating voltage and current. I do not have to tell you who won. Alternating current has great advantages over direct current, as I will show you below.

A sinusoidal current is just that, Figure 6.8. The voltage we get at sockets in the USA is rated as 120 V, 60 Hz, AC (AC for alternating current) but the actual voltage as a function of time goes from 0 to 170 V back to 0 and down to −170 V, that is, the same voltage but with the polarity reversed and it goes up and down 60 times a second (in most of Europe the voltage is 220 V and 50 Hz). The reason that a sinusoidal voltage that goes from 170 V to −170 V is rated at 120 V is because the power dissipated across a resistor is

the same as it would be for a DC (direct current) of 120 V. The time between peaks is 0.01667 s, that is 1/60 s.

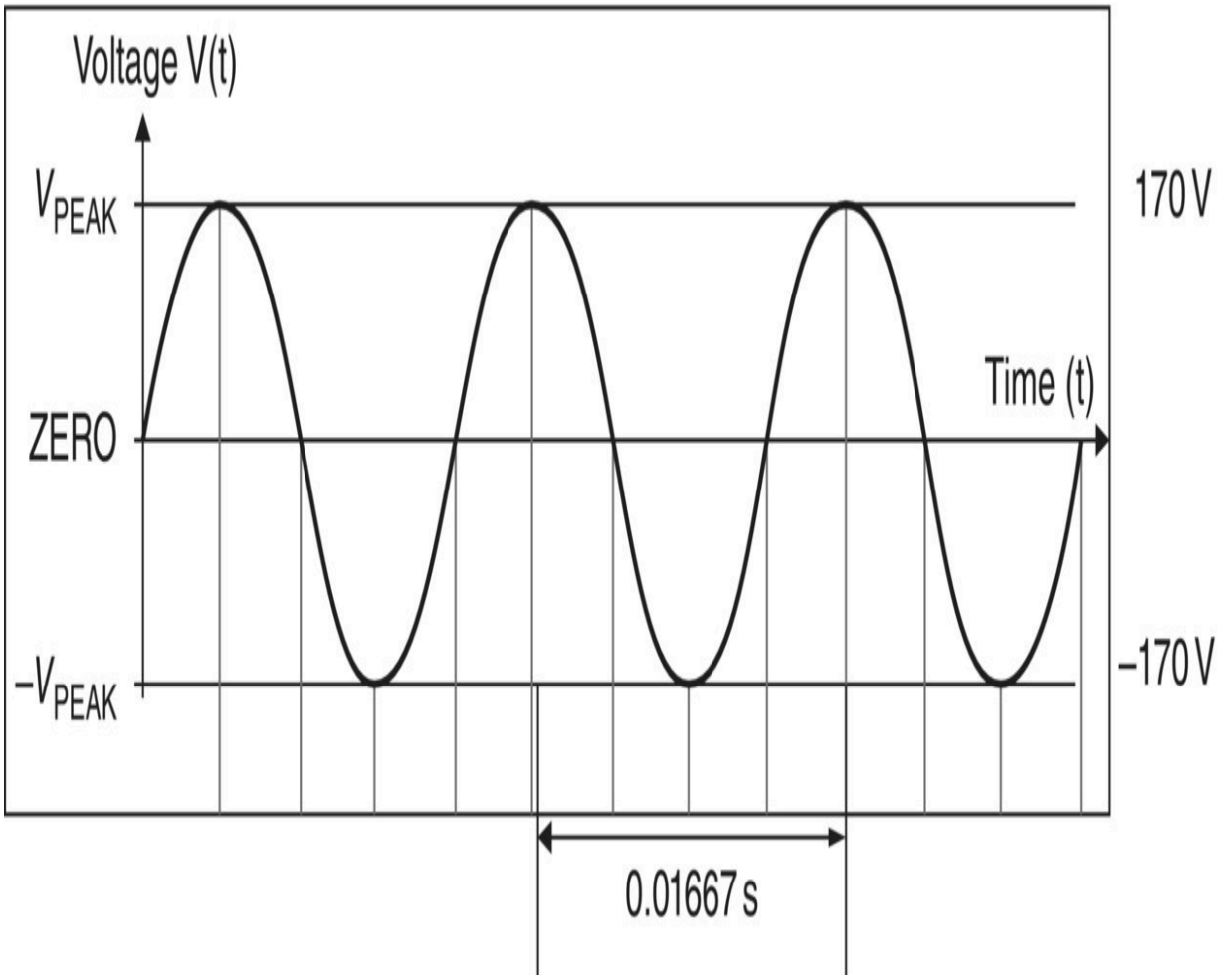


Figure 6.8 The 120 V electrical oscillating voltage, AC, in the USA.

Electrical systems and large appliances use almost uniquely AC voltages. Microelectronics use almost uniquely DC voltages. That is why most, if not all, electronic devices need a power cord with a rectifier, either in the line or in the device, that changes the AC to a low voltage DC. I will talk a about these rectifiers in the next chapter.

Sinusoidal voltages and currents change the way that capacitors and inductors affect the currents. A sinusoidal voltage across the capacitor is constantly sending electrons to one metal plate and removing it from the other and vice versa. If you imagine the

capacitor as a room with two doors and you see people similarly dressed coming in through one door and leaving through another and then back again you won't know that in the middle of the room, there is a wall dividing the room in half. For all practical purposes the two doors are connected. I will explain the effective resistance that the capacitor presents to a sinusoidal voltage in Appendix 6.1.

The same thing with the inductor. If the current keeps on changing, up and down, the magnetic field will increase, then decrease to zero and change directions constantly. Also, it will oppose the flow of the current and I quantify that resistance in Appendix 6.1.

6.6 Inductor Applications

Since I will not discuss inductors in the rest of the book, I'd like to mention here a couple of its applications.

If you drive a car, you go over inductors all the time. They consist of four to six turns of wire inserted in the pavement. This coil is placed about 5 cm under the pavement and the diameter of the coil's loop is roughly 1.5 m. There is always a small AC voltage applied to the coil. The inductance presents a resistance to the current. When a car, truck, bicycle or large piece of metal goes over it, the inductance of the coil changes and therefore the current also changes. Because the change in the current depends on the specific metallic object on top of the coil and it depends on the type, size, and height of the vehicle, it is possible to identify not only the presence of a car but what type of car is going over it.

One of the most common applications of the inductor and the reason Westinghouse won over Edison, is the transformer, which works only with sinusoidal voltages Figure 6.9.

We apply a sinusoidal voltage at the primary coil, and this generates a magnetic field which, as we have seen, is proportional to the number of turns of the coil. Since the AC current changes continuously from positive to negative and back again, the magnetic field also constantly changes in magnitude and in direction, pointing

up when the current goes in one direction and pointing down when it goes in the opposite direction. A magnetic core directs this magnetic field to the secondary coil. The magnetic field is the same in both coils. The voltage of the secondary coil is proportional to the number of turns in its coil. So, if there are twice as many turns in the secondary coil as there are in the primary coil, the voltage across the secondary coil is twice as high. But because the power has to be the same (you cannot create power out of nothing), the current is half as much as the current in the primary coil.

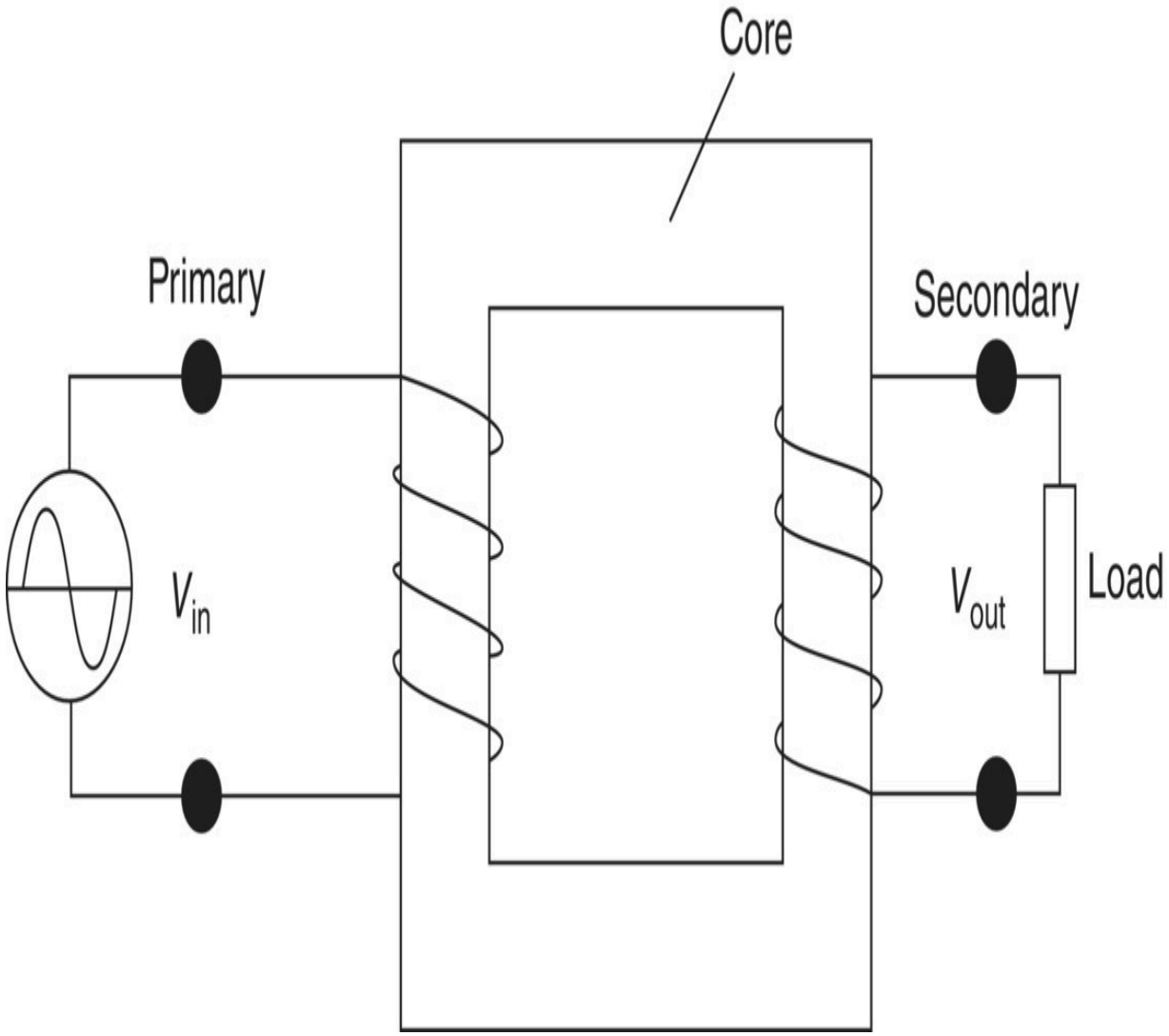


Figure 6.9 A transformer consists of two coils sharing the same magnetic core and thus sharing the same magnetic flux.

We can use this property to generate and send electricity from one point to another and change the sinusoidal voltage to whatever is appropriate for our use. Figure 6.10 shows a sketch of a typical electrical distribution system.

A very large hydropower generator can output 1000 MW (megawatts) with a voltage of 20 kV (kilovolts). A transformer increases the voltage from 20 kV to 200 kV. The transformer between the generator and the transmission line has a ratio of 1 to 10. As we go to the substation, the voltage goes back to 20 kV with another

transformer with a ratio now of 10 to 1. We proceed to a factory that may need a high voltage of 10 kV. So, another transformer with a ratio of 2 to 1 performs that change. As we go to the houses, we need to bring down the voltage to 120 V. So, a final transformer with a ratio of 167 to 1 provides the right electricity to our outlets.

Why go to all this trouble? Well, you already know the answer. We have seen that the power dissipated in a resistor is equal to the square of the current times the resistance, $P = I^2R$, Eq. (6.3). The resistance of a line is quite low, but not zero. For a given power, if I raise the voltage by a factor of 10, the current decreases by the same amount ($W = V \times I$), and the power lost in the resistance of the transmission line, as small as it may be, it is now 100 times lower ($P = I^2 \times R$). If I were to increase the voltage by a factor of 20, the power dissipated in the line would be 400 times less. (This is a simplification. There are other electromagnetic and skin effects that limit how high we can go, but you get the idea.) To save power in the transmission we want the voltage to be as high as possible, so that the current is as small as possible. There are limitations. We don't want the voltage so high that sparks flow from one line to another or to the person walking under the lines. Nothing is as simple as a simple theory would tell you.

Now that we have an idea of how these electrical/electronic components work, we are ready to take a look at many of the uses of semiconductor pn-junctions, the diodes.

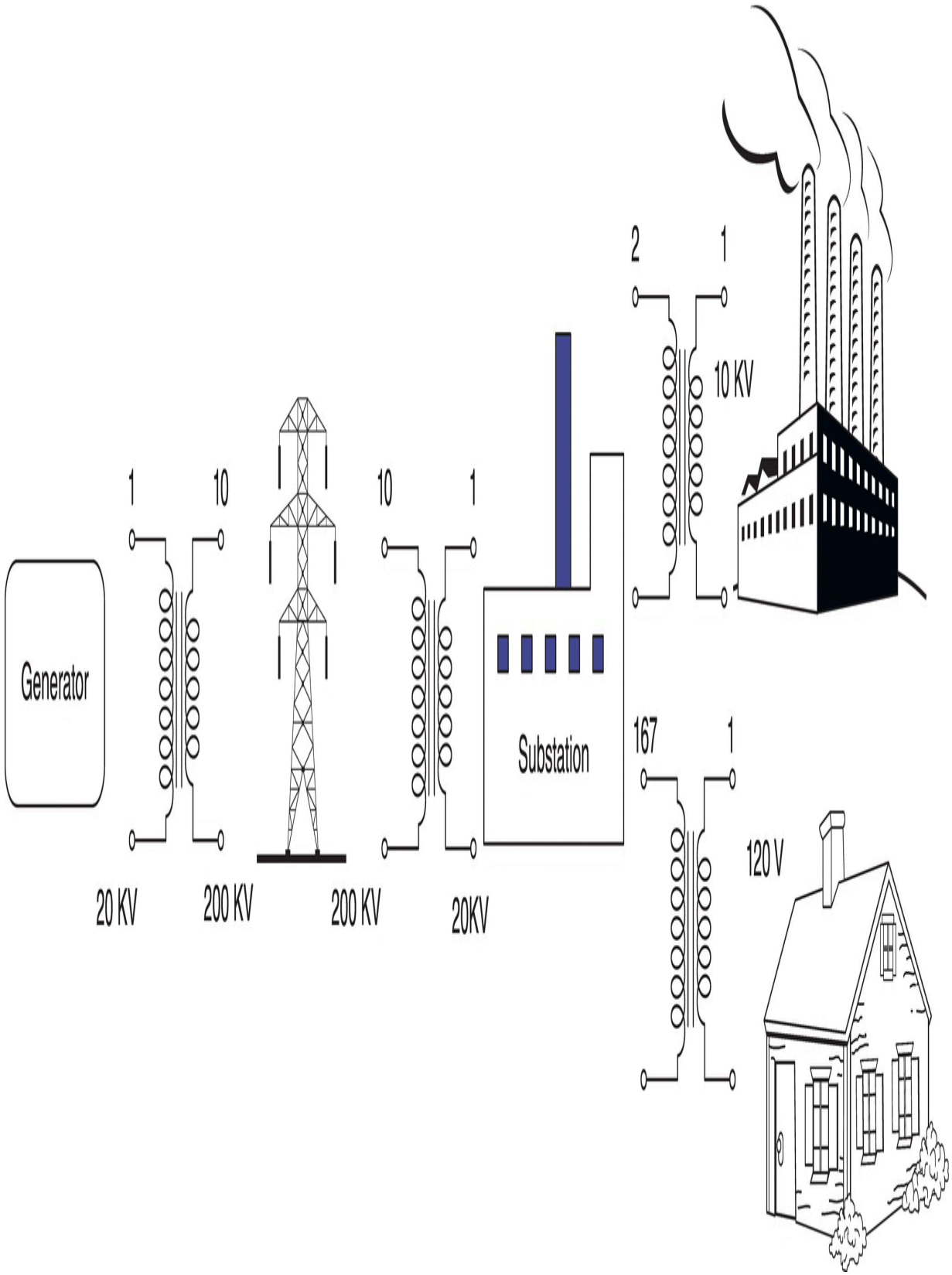

Generator
1
10
20 KV
200 KV
200 KV
10
1
20KV
Substation
2
1
10 KV
167
1
120 V

Figure 6.10 Using transformers in an electrical distribution system we can efficiently transport electricity from the source to the user.

6.7 Summary and Conclusions

After clarifying a little the difference between voltage and current we have studied what we call the passive elements and the relationship between voltage, current, and power in these components. In the resistor the relationship is given by Ohms law, $V = IR$ and the power by $P = V \times I = I^2R$. We have seen that the capacitance can store charges until the voltage across its plate is equal to the applied voltage. The current through a capacitor is proportional not to the voltage but to the change of the voltage. The inductor stores energy by the magnetic field it generates. In the inductor the voltage changes as a function not of the current but of the change of the current.

The value of all of these elements can be calculated by knowing the property of the materials we use to make them: resistivity for the resistor, permittivity for the capacitor, and permeability for the inductor.

With the understanding of these components we are ready now to explain the many practical circuits we can design with pn-junction devices, that is, diodes.

Appendix 6.1 Impedance and Phase Changes

In the previous sections I covered the capacitor and inductor qualitatively. Here I will explain a little more the performance of these two devices under a sinusoidal input voltage. Look at Figure 6.11.

First, compare the top two sinusoidal waves and pay attention to the two vertical dashed lines. The top wave is the source voltage and the one below is the current through the resistor. Both waves are in-phase, that is, when the voltage goes up, the current through the

resistor also goes up and when the voltage goes down to zero, the current is also zero. That is, except for the units and the magnitudes, the two curves are identical and they fit one on top of the other.

Now let's compare the voltage wave with the current through the capacitor. There is current through the capacitor only when the voltage changes, Eq. (6.16). When the voltage is at the top, the voltage does not change, or more precisely the instantaneous voltage just before it hits the top and the instantaneous voltage just after are the same. So, the voltage at that instance of time does not change and therefore the current through the capacitor is zero (follow the dashed line and compare the first to the third sinusoidal waveforms). That is what the dashed line on the left shows. As soon as the voltage starts decreasing, the charges stored in the capacitor start moving back and when the instantaneous voltage reaches zero (dashed line on the right), that is, when the voltage changes most rapidly, the current through the capacitor is largest in the negative direction. The sinusoidal current is shifted 90° ahead of the voltage.

The opposite is the case for the inductor. For clarity, let's compare the wave of the current through the inductor (the lowest wave) to the voltage (the uppermost wave) instead of the other way around. Now there is voltage only when the current changes, Eq. (6.18). First, taking a look at the left vertical dashed line, we see that when the current is zero and changing rapidly from negative to positive the voltage is highest and when the current is highest (dashed line on the right) and therefore not changing, the voltage is zero. The current wave is ahead of the voltage wave by 90°.

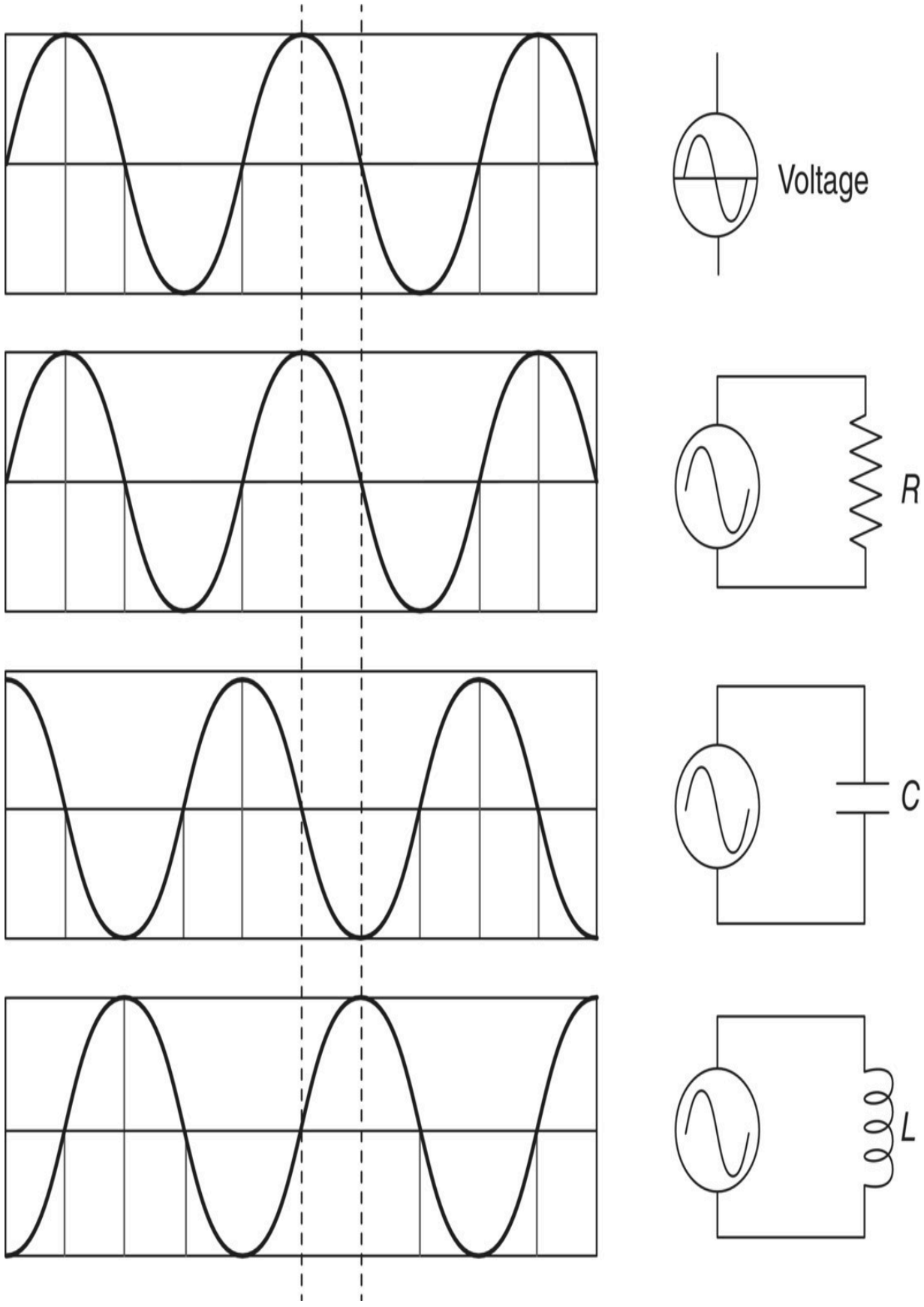

Voltage
R
C
L

Figure 6.11 The sinusoidal current through a resistor is in phase with the voltage, through a capacitor the current is ahead of the voltage, and through an inductor the current is lagging behind the voltage.

A sinusoidal voltage can be written mathematically as:

$$v(t) = V\sin(2\pi ft + \phi) = V\sin(\omega t + \phi) \tag{6.19}$$

where ϕ is the phase shift and ω is just the product $2\pi f$. I now use small v to indicate that the voltage is no longer a constant but a function of time. When the voltage is changing constantly, there is a current "through" the capacitor and inductor. I put quotation marks around the word "through" because in the capacitor no electrons physically move from one plate to the other, but electrons and positive charges are being added and removed periodically from each plate, thus from the outside it looks like a current is passing through it.

These two devices, like the resistor, also oppose the flow of current. We call this *reactance* and is has the symbol capital X. The reactance is given by

$$X_C = \frac{1}{\omega C} \quad \text{and} \quad X_L = \omega L \tag{6.20}$$

This makes sense, since, when the frequency goes to zero, that is, we have a constant DC, voltage or current, ω goes to zero and the capacitive reactance X_C goes to infinity, which is what you would expect if we have a DC voltage connected to the capacitor. The inductive reactance, X_L, goes to zero with no opposition whatsoever under a DC condition. The values of the reactance above are just the magnitude of the opposition to the current. There is also the phase shift.

When we have a combination of any or both of these two devices with the resistor, the total opposition to the current, we call it

impedance, with symbol *Z*, is:

$$Z = \sqrt{R^2 + \left(X_L - X_C\right)^2} \tag{6.21}$$

and the phase is no longer 90° but a combination depending on the different values of the resistance and reactance. We can calculate the new phase shift by

$$\phi = \tan^{-1} \frac{X_L - X_C}{R} \tag{6.22}$$

7
Diode Applications

OBJECTIVES OF THIS CHAPTER

In Chapter 5 I covered the simplest, most basic device we can make by combining an n-type semiconductor with a p-type semiconductor. The pn-junction allows current to flow in just one direction. That, by definition, is what a diode does. This property is extremely useful. So now that we know how the pn-junction works and the electrical properties of the passive elements, resistors, inductors, and capacitors, we are ready to discuss the large number of very useful applications of diodes, from solar cells to rectifiers to a variety of other uses. In this chapter I discuss all of these devices using only pn-junctions, including solar cells, rectifiers, voltage doublers, surge protectors, and others.

7.1 Solar Cells

One of the most important and valuable applications of the pn-junction is the solar cell. The solar cell is a very simple application of a diode. It does not need a complicated circuit to understand how it works, just the basic properties of the pn-junction (see Figure 7.1).

The top figure shows the standard structure of a semiconductor diode, an n-type material adjoined to a p-type semiconductor. In the middle we have the depletion or the transition region. This is the region where electrons from the n-type material diffuse toward the p-type material and in doing so the n-type material becomes more positive until the internal potential cancels the diffusion forces. The

lower sketch shows the electrostatic potential inside the diode. We have seen this already in Chapter 5, Figure 5.3, where the n-type material gets more positive and the p-type more negative. Now suppose that a photon, that is, a quantum of light, hits an atom in the depletion region. The photon transfers its energy to the atom. The energy frees an electron, creating an *electron–hole* pair as I show in the lower part of Figure 7.1. Since there is a potential gradient, the electron moves to the left, toward the positive potential, and the hole moves to the right, toward the negative potential. Note that this motion of electrons and holes happens without the need of any battery or any external source: it is the internal field of the pn-junction itself that separates and moves the electrons and holes in different directions. As a matter of fact, now the diode becomes an electrical "generator." See how simple it is? We call this a photovoltaic (PV) cell. This simple device may save the world from climate explosion.

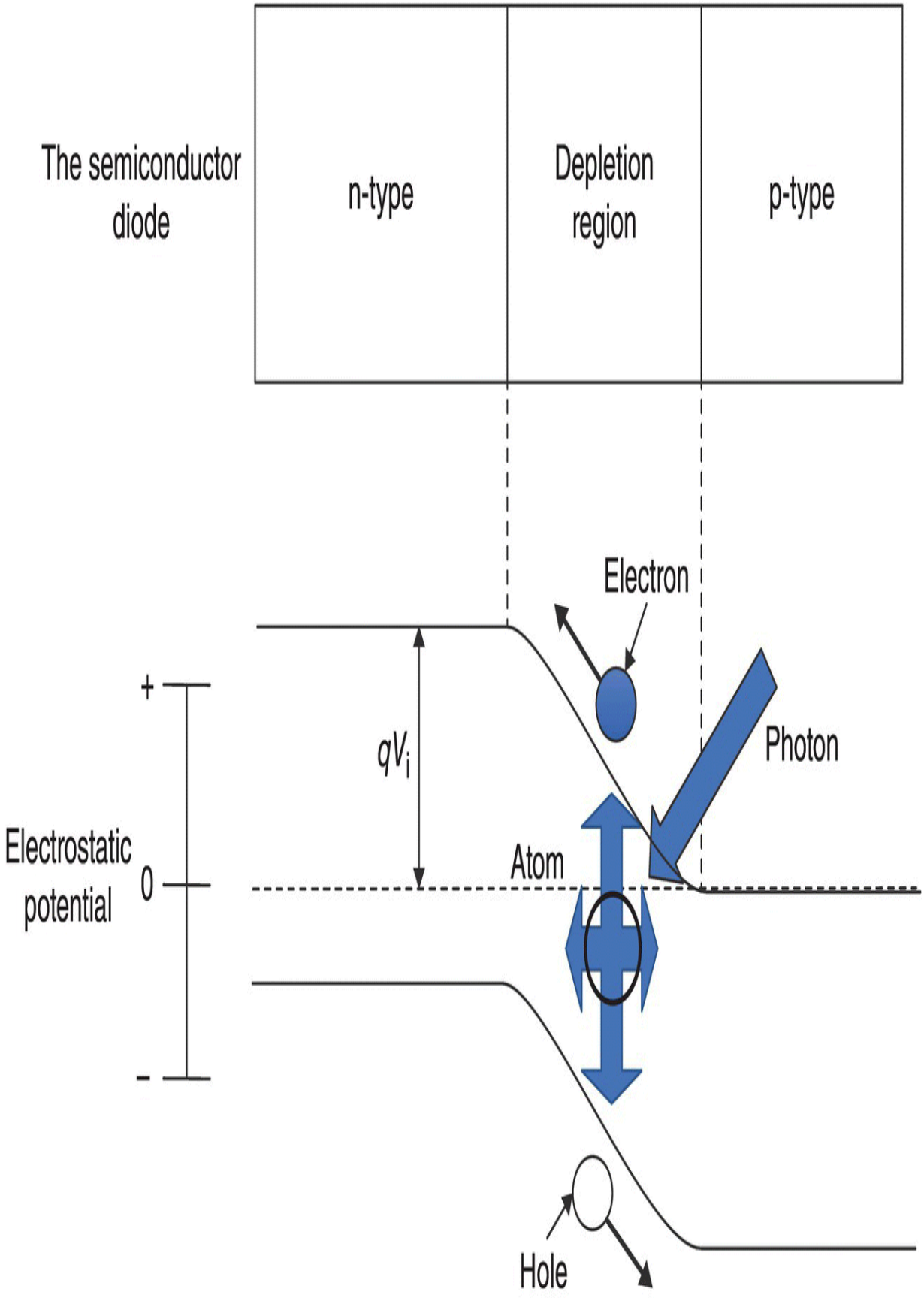

The semiconductor diode
n-type
Depletion region
p-type
Electron
Photon
Atom
qV_i
+
0
-
Electrostatic potential
Hole

Figure 7.1 When a photon strikes the transition region of a solar cell the energy released creates an electron–hole pair, the electron moving to the positive and the hole to the negative potential.

The sun gives out about 500–1000 W of energy per meter square per hour. If you assume an average roof area of 100 m^2, covered with solar cell panels, and maybe six hours of sun, we get 300 kW a day. The efficiency of solar cells, especially those that are most commonly used, and thus less expensive, is only about 15%, so we collect about 45 kW a day. A typical house uses about 25 kW a day.

The main semiconductors we use for PV cells are silicon, gallium arsenide (GaAs), cadmium telluride (CdTe), and cadmium sulfide (CdS). GaAs is the most expensive, but it has the best efficiency.

7.2 Rectifiers

By far the main application of diodes in electronics is as rectifiers. As I mentioned before, electronic devices use DC (direct current). We need a rectifier to change the sinusoidal input voltage to a constant voltage (Figure 7.2).

When the sinusoidal input voltage, V_{in}, is positive, the current goes through the diode with no difficulty and the entire voltage drops across the resistor R. The output voltage in this circuit, V_{out}, is for all practical purposes equal to V_{in} during the positive input cycle. When the input sinusoidal voltage, $V_{in,}$ becomes negative, the diode opposes the flow of the current so the voltage across the resistance is zero. (Actually the voltage is very small because of the tiny reversed biased current. For all practical purposes the current through the resistor and therefore the voltage across the resistor is insignificant.)

That is the first step of rectification. A voltage that goes back and forth from positive to negative, now is only positive. That is not yet a real direct, constant, voltage like that of a battery, but it is the first

step. The next step is to smooth out the voltage to make it more constant. We do this by adding a capacitor in the circuit (Figure 7.3).

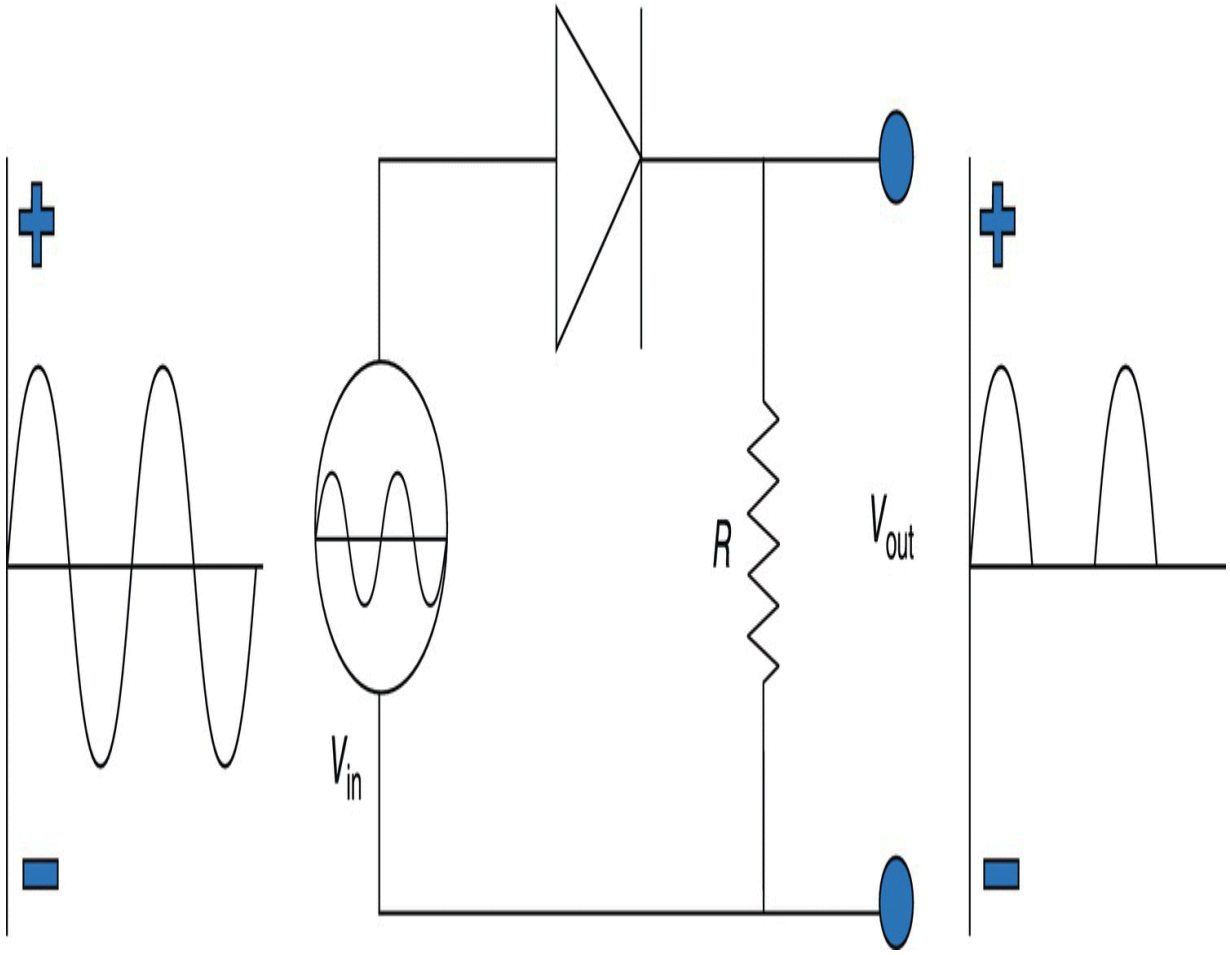


Figure 7.2 A rectifier circuit only lets the positive swing of the current pass through the diode.

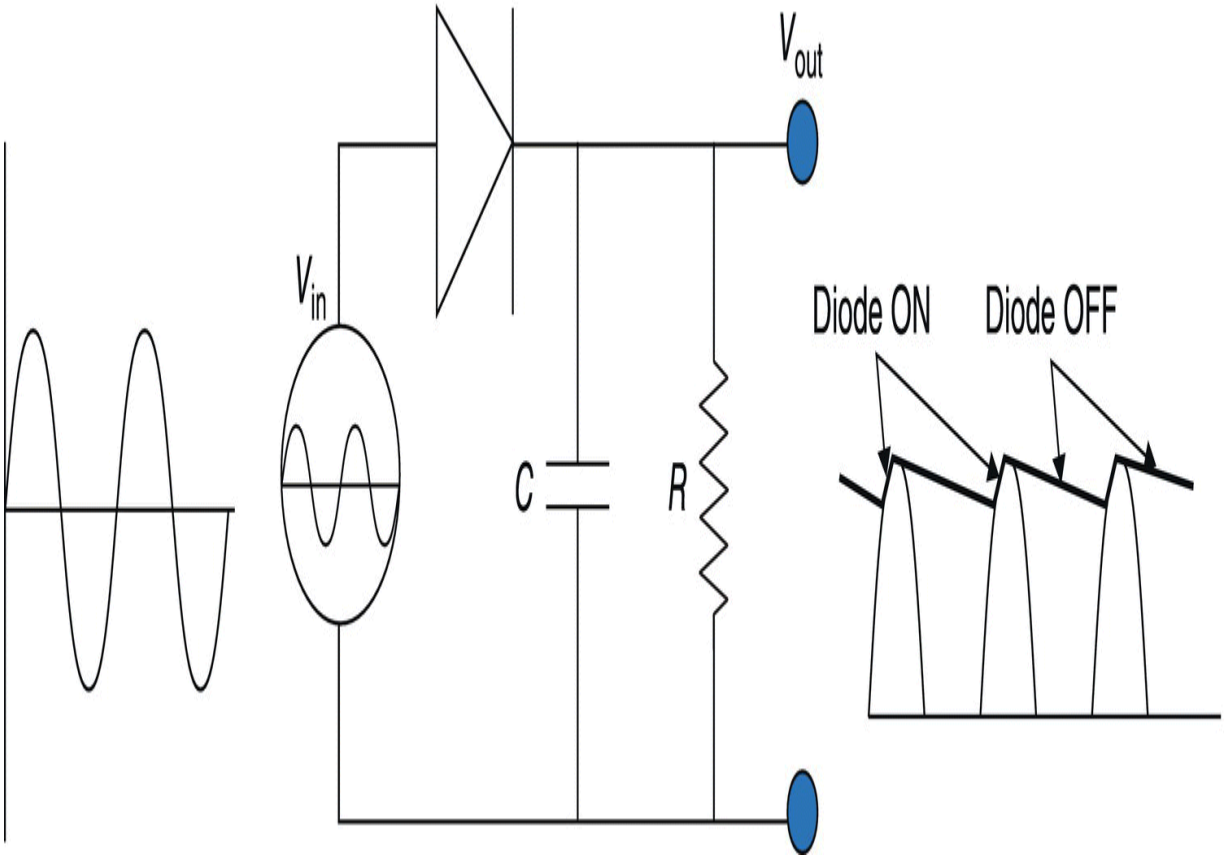


Figure 7.3 A capacitor in parallel with a resistor stores charges during the positive cycle and releases charge during the negative cycle, smoothing the output voltage.

Now look at what happens if we add a capacitor in parallel with the resistor. When the sinusoidal input voltage is positive, the current flows through the diode and in addition to flowing through the resistor R, it increases the voltage across the capacitor C to the maximum positive input voltage. As the sinusoidal input voltage starts decreasing, the output voltage, V_{out}, stored in the capacitor, is now higher than the input voltage, V_{in}. At this point no more current flows through the diode, which is now reversed biased. The electrons that have instantaneously charged the capacitor start discharging through the resistor. Depending on the value of the capacitance and the resistance, the next positive input voltage peak will hit the diode before the capacitor has had time to completely discharge. At that point, the diode becomes forward biased again,

current flows and the capacitor voltage goes back to the maximum positive input voltage. The output voltage is not yet really constant but is getting there. How smooth the output voltage is depends on what the value of the capacitor is compared to the resistor. Appendix 7.1 explains in more detail how to select a resistor and capacitor combination to obtain the desired constant output voltage.

Of course, we can do better than that. Figure 7.4 shows a rectifying circuit which uses both the positive and negative swings of the sinusoidal input voltage to charge the capacitor to the maximum positive voltage.

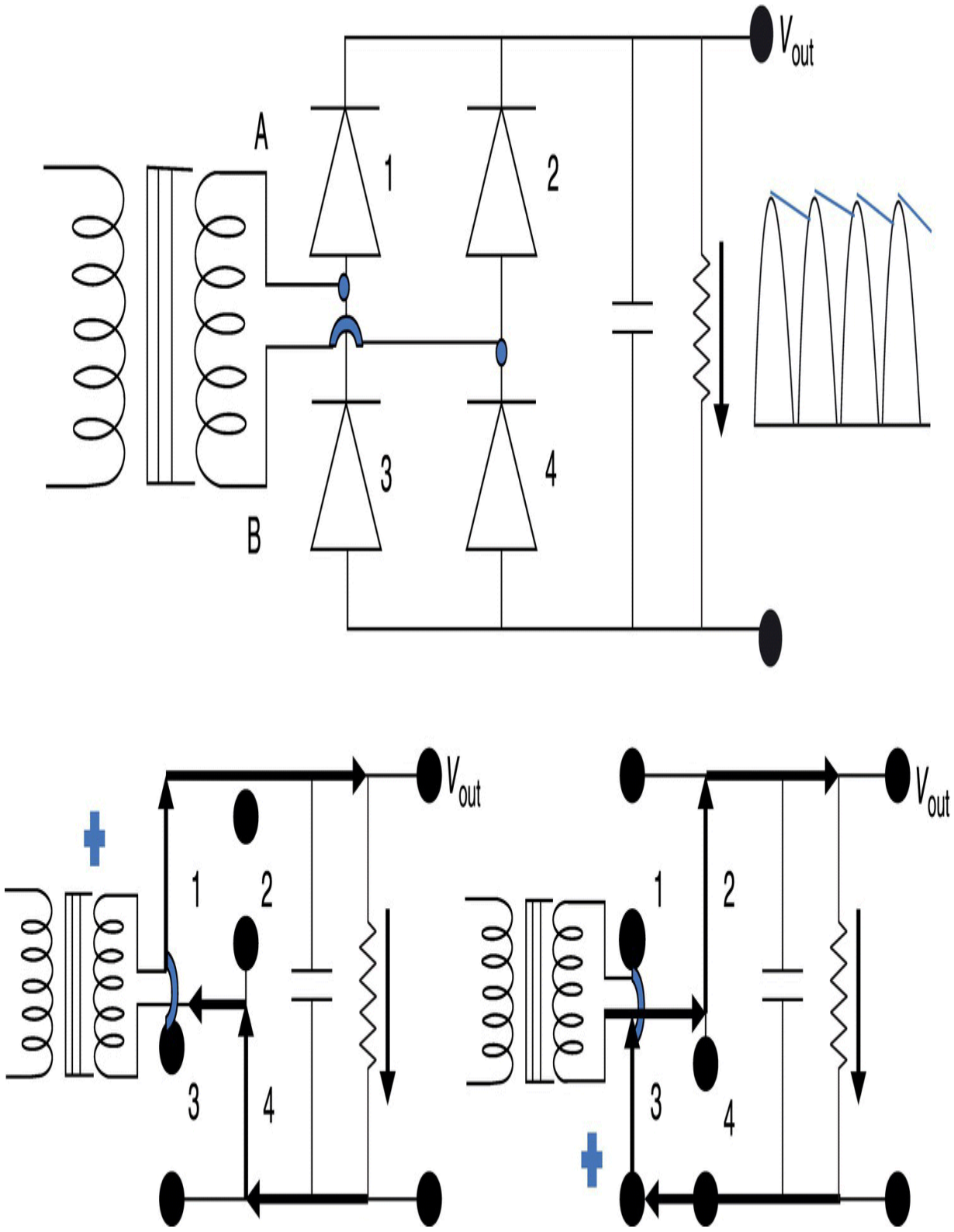


Figure 7.4 A full-wave rectifier with a smoothing capacitor uses both positive and negative swings of the input voltage.

I have added now a transformer. Most electronic devices work with DC voltages between 2 and 15 V. iPads, for example use 5 V. The transformer brings the sinusoidal wave amplitude down from 120 V to 5 V. Now instead of one diode, I use four (top of Figure 7.4). When the input voltage at point A is positive, and therefore point B is negative, diodes 1 and 4 are forward biased and 2 and 3 reversed biased. (Note the small bridge drawn on the line between diode 4 and terminal B. This is the standard electronic symbol indicating that the horizontal line is *not* connected to the vertical line between diodes 3 and 1.) The current flows from terminal A to diode 1, charges positively the upper plate of the capacitor to the maximum voltage A, flows down through the resistor and the current returns to point B through diode 4. Note that in this half cycle diodes 2 and 3 are reversed biased so they do not conduct. This is the situation I show in the sketch at the lower left of Figure 7.4.

At the negative half of the sinusoidal cycle, the negative cycle, point A is negative and point B is positive. Point B, being positive, makes diodes 2 and 3 forward biased and 1 and 4 reversed biased so now diodes 2 and 3 are conductive. The positive current goes from B, through diode 2 and, voilà, also down the resistor, returning to point A through diode 3. I show this case at the bottom right of Figure 7.4. The current through the resistor goes down during both cycles of the input voltage, the positive and negative cycles. Now the capacitor is kicked up to the maximum input voltage twice as often as in the circuit of Figure 7.2, in the US electrical system 120 times a second instead of 60.

There are other ways you can get full rectification. Figure 7.5 shows a different way. In this case the output voltage is half the magnitude of the voltage between A and B. Can you figure out how it works?

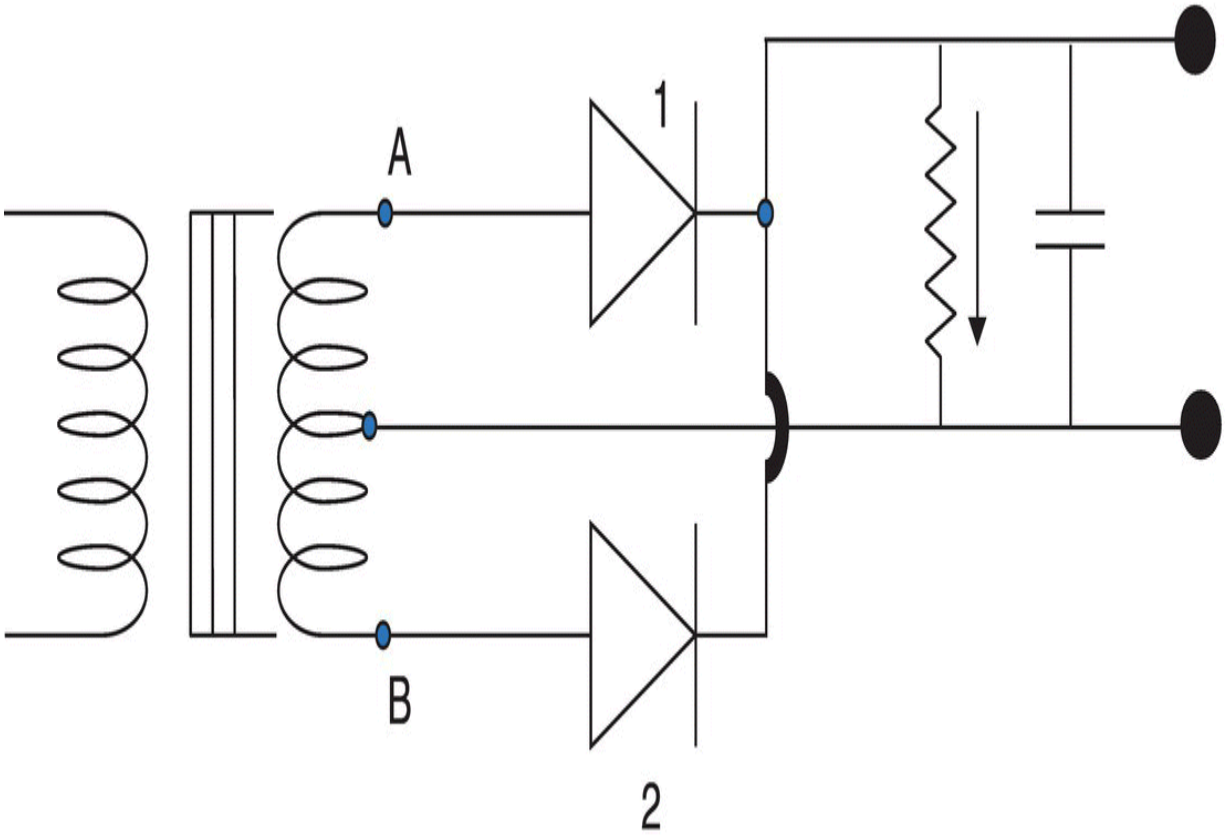


Figure 7.5 Full-wave rectification using the middle tap of a transformer.

7.3 Current Protection Circuit

Another common use of a diode is as a reversed current protection circuit. Many delicate electronic devices would be destroyed if you were to install the batteries backwards (how many times have I done that!). A very simple circuit protects the device from careless battery changing. Figure 7.6 shows one such circuit. If the batteries are reversed, the diode blocks the current, preventing any damage. The dotted box represents a delicate circuit that has some equivalent input resistance.

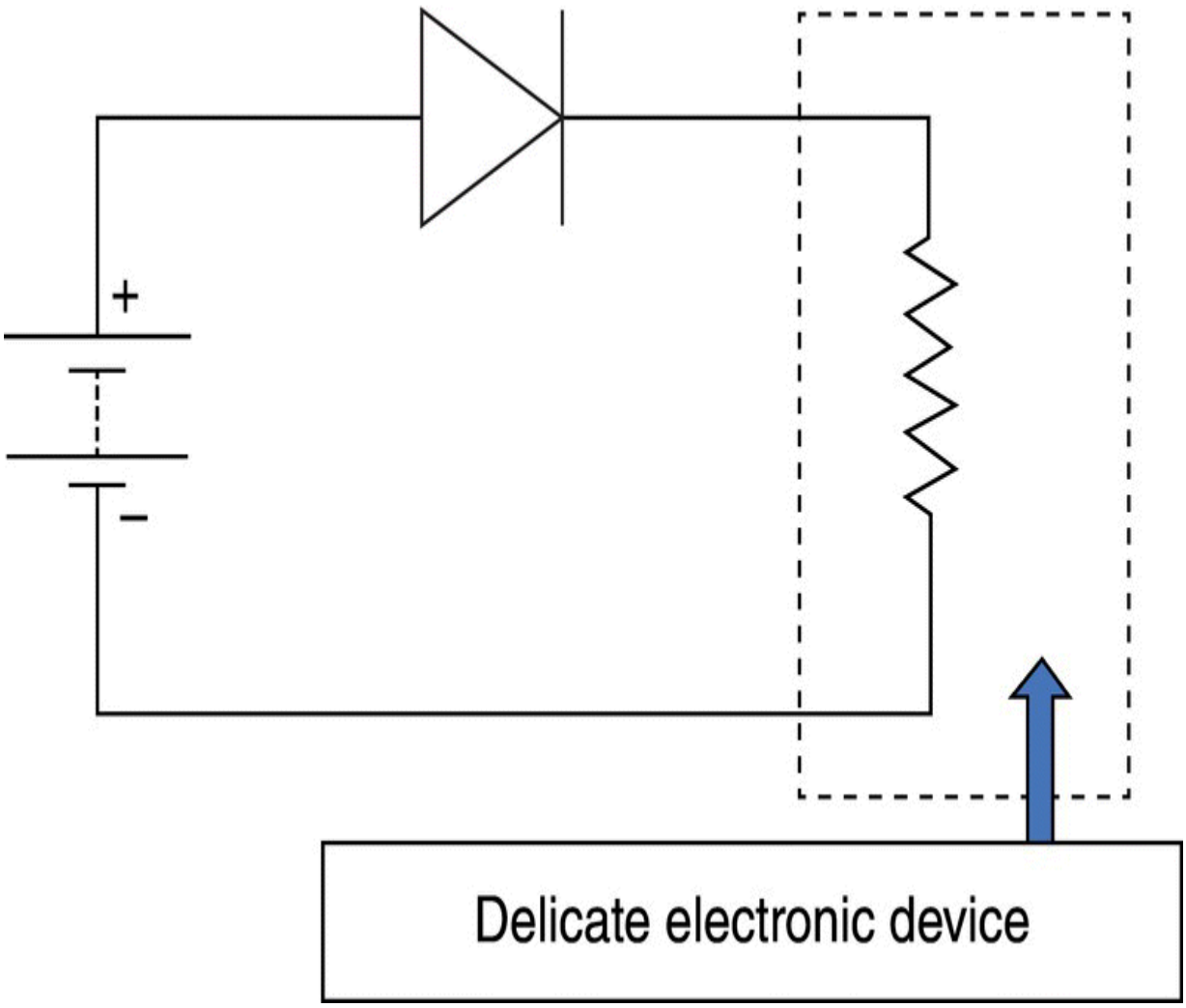


Figure 7.6 A reverse current protection circuit prevents damage to the delicate electronic circuit if the batteries are reversed.

7.4 Clamping Circuit

Another application of diodes is as clampers. A clamper shifts the zero value of a sinusoidal wave so that the entire sinusoidal function is positive (or negative if I reverse the diode). The positive peak of the sinusoidal voltage is twice as high, and the most negative value is zero (or very close to it), that is, the sinusoidal output voltage instead of going from −10 to +10 V, now goes from 0 to +20 V. Figure 7.7 shows the circuit.

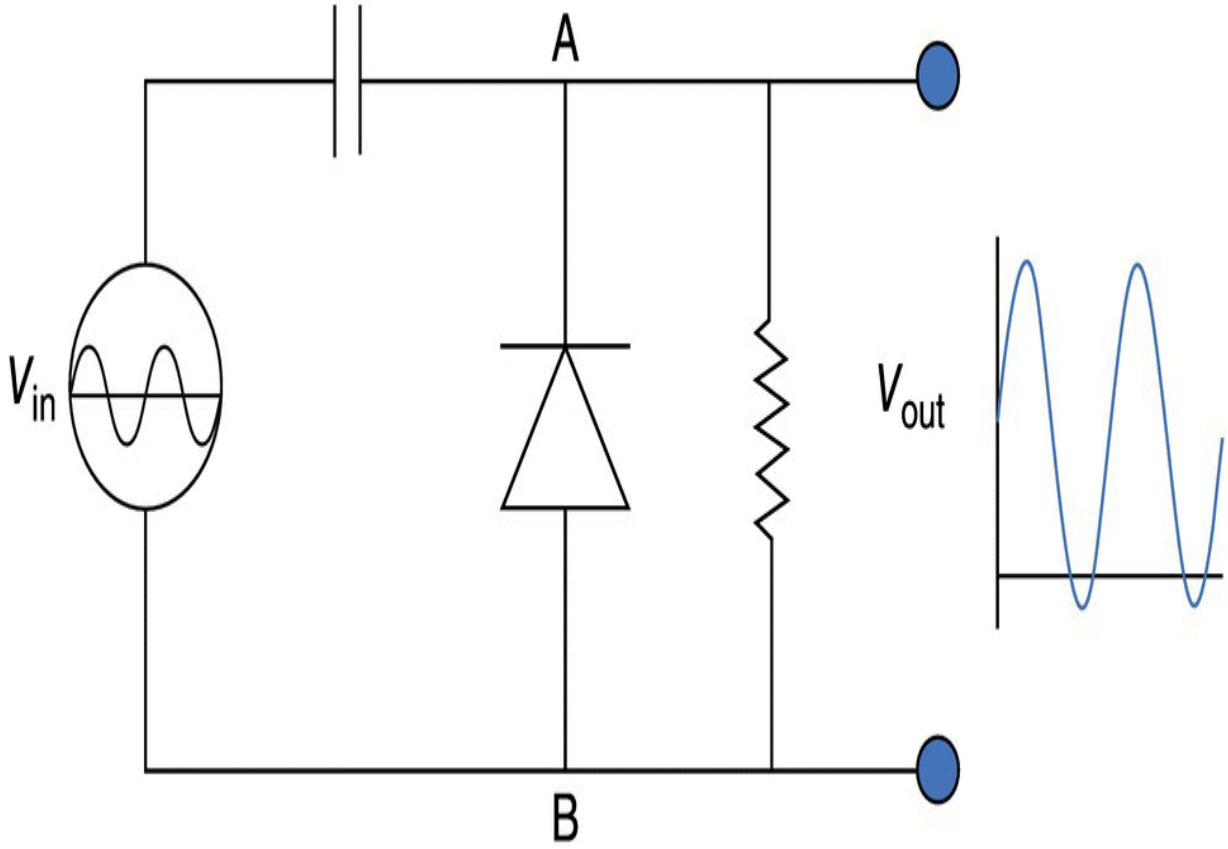


Figure 7.7 A clamping circuit shifts the sinusoidal wave so that the entire sinusoidal wave is positive.

When the input voltage, V_{in}, is in the negative cycle, point B is positive, the diode is forward biased and therefore current flows through it and charges the capacitor so point A becomes positively charged to the maximum value of the negative cycle of the sinusoidal voltage. Let's use some numbers. Suppose that the input voltage, V_{in}, goes from −10 to +10 V. When the input V_{in} is in the negative cycle, point B is positive, +10 V, the diode acts as a short circuit and the current charges the capacitor, point A, to +10 V. As soon as the input voltage starts decreasing, point A becomes more positive than point B, and the diode becomes reversed biased and does not conduct, it is open, so now, at its maximum positive value of V_{in}, the voltage at point A is the sum of the value V_{in} plus the voltage across the capacitor, so point A is now +20 V. When V_{in} becomes negative again, the voltage at A is 0 (10 V from the

capacitor minus 10 V from the negative swing). The voltage output V_{out} has shifted, as I show on the right of Figure 7.7.

I should make a clarification which I did not mentioned before so as not to complicate matters. Remember that the diode has a turn-on voltage between 0.5 and 0.7 V. Therefore, the voltage output is not really between 0 and 20 V, but something between −0.5 and +19.5 V. This effect applies to all the diode circuits I describe in this chapter.

7.5 Voltage Clipper

In many cases we want to be sure that the voltage does not exceed a certain value. We may have an input voltage, let us say 10 V, but parts of the circuit should not run higher than 5 V. That is when we use a voltage clipper (Figure 7.8), a simple circuit that accomplishes this task.

Suppose, for example, that the sinusoidal input voltage goes from +10 to −10 V and we want the output to be limited to 5 V positive and negative. I would connect both batteries of 4.5 V (remember that the diode has a turn-on voltage of 0.5 V) with the polarity I show in Figure 7.8. When the sinusoidal input voltage, V_{in}, starts going up in its positive cycle, both diodes are reversed bias, no current goes through either one of them, thus the voltage output, V_{out}, across the resistor is equal to V_{in}. When V_{in} exceeds the voltage at V_1 (plus the turn-on voltage of the diode), D1 turns on, it shorts the circuit, and keeps the output voltage constant at 5 V, the voltage of battery V_1. Similarly, in the negative cycle, D2 shorts as soon as the voltage increases above −5 V. Both diodes are off when the input voltage is less than the voltages of V_1 and V_2. Notice that the clipping does not have to be the same in both directions. It is up to me to limit the current any way I want to by choosing different batteries.

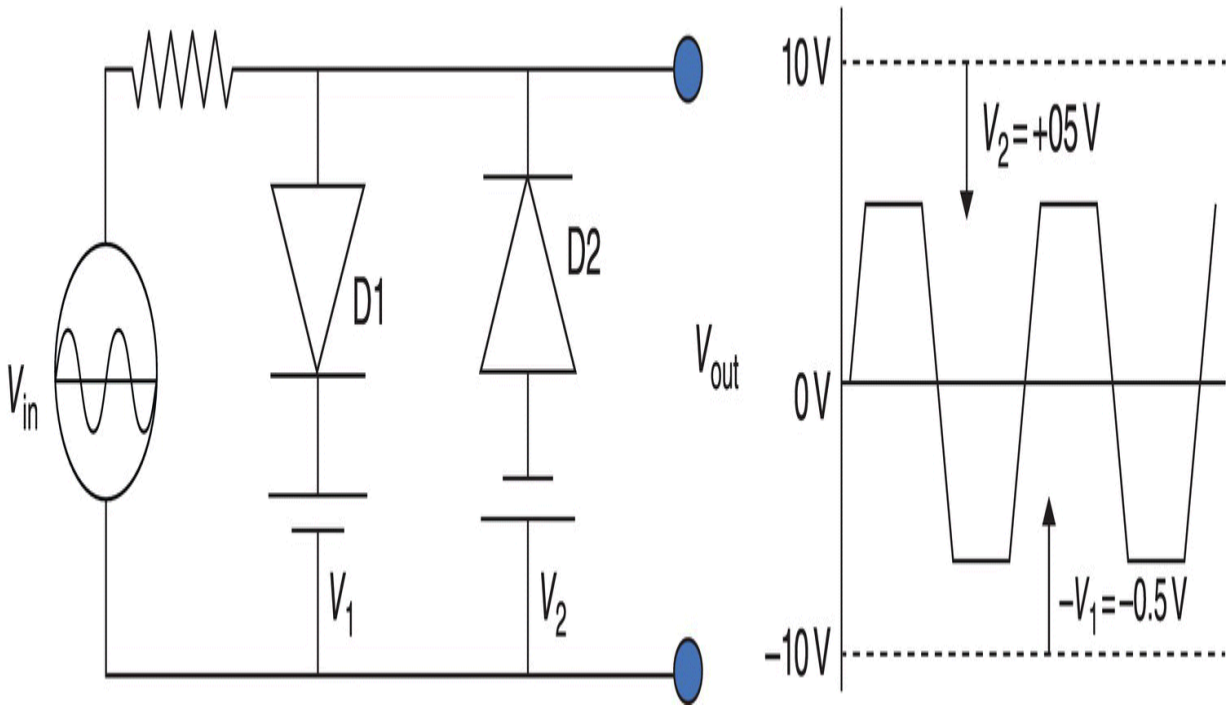


Figure 7.8 A voltage clipper prevents the output voltage going over a specified value.

7.6 Half-wave Voltage Doubler

Ready for one final circuit? The voltage doubler (Figure 7.9).

First notice that point B is connected to the lower terminal of the input voltage and one terminal of diode D1 and capacitor C2, and the lower output voltage terminal. Thus, this lower line is always at the same potential as the lower input voltage contact. Let us assume again that the input voltage is a 10 V sinusoidal voltage.

To understand what happens, let me make a simplified model of the circuit when the input voltage is positive and when it is negative (Figure 7.10).

Take a look first at the upper drawing in Figure 7.10. In the negative input voltage cycle, line B is positive, therefore diode D1 is conducting. Diode D2 is reversed biased, thus it is open and no current flows through capacitor C2. Capacitor C1 charges until point

C is at +10 V, the same as point B. Since diode D2 is open, capacitor C1 cannot discharge.

As soon as the sinusoidal input voltage starts increasing from its most negative value, point C, stuck at +10 V, is more positive than line B and thus diode D1 turns off (lower diagram in Figure 7.10). Now D1 is open and D2 is conducting. Point A is now +10 V and the voltage across C1 is also +10 V so points C and D, which are now connected, are both at +20 V.

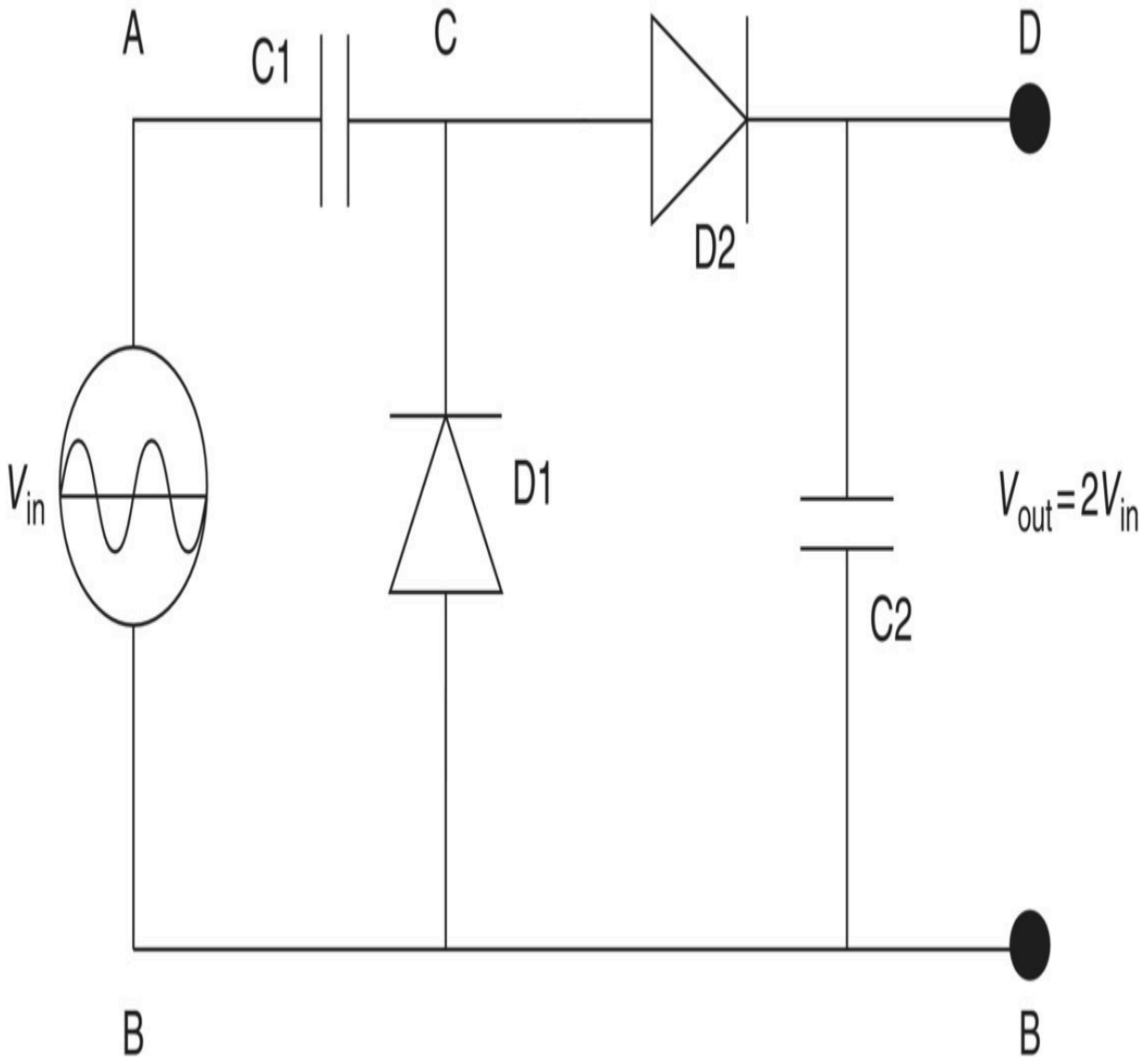
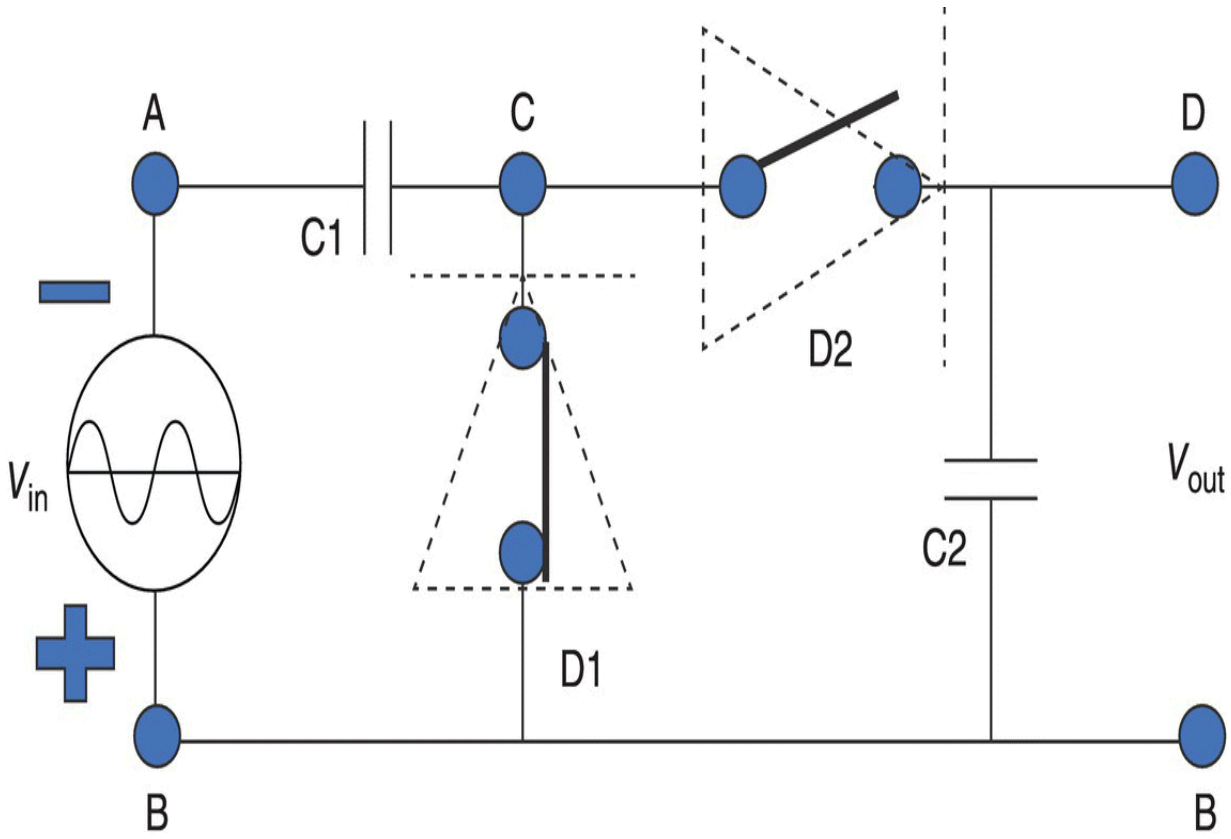


Figure 7.9 A half-wave voltage doubler circuit results in an output voltage twice as high as the input voltage.


A
C
D
C1
D2
V_{in}
V_{out}
C2
D1
B
B

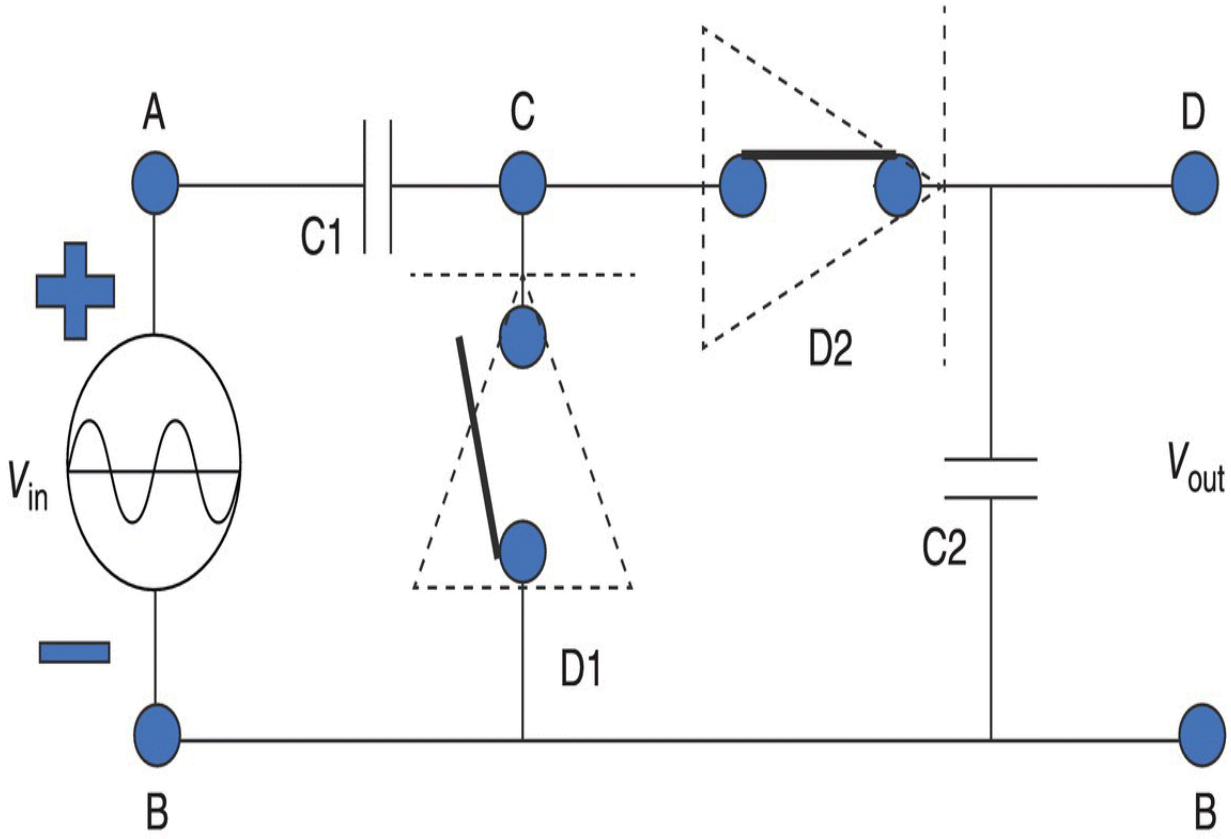

A
C
D
C1
D2
V_{in}
V_{out}
C2
D1
B
B

Figure 7.10 A simplified equivalent circuit for a voltage doubler.

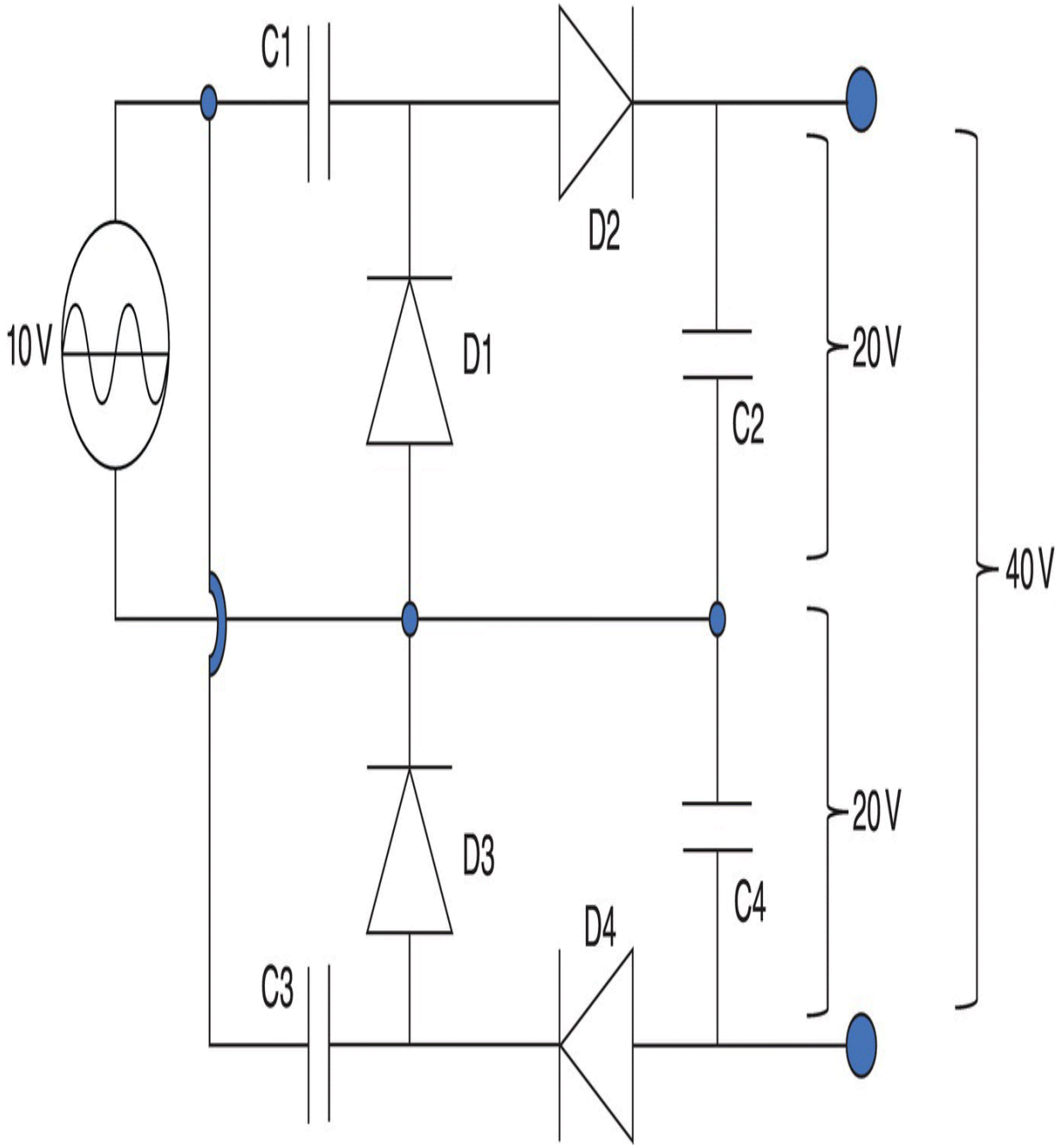


Figure 7.11 A circuit that makes the output voltage four times as high as the input voltage.

What we have accomplished with this circuit is to generate a rectified full voltage twice as high as the maximum input sinusoidal voltage. What I explain above is the steady-state condition.

Could I increase the output voltage still more than just doubling it? Yes, Figure 7.11 shows a circuit where the output voltage is four times that of the input voltage. No explanation is needed. You can figure it out.

7.7 Solar Cells Bypass Diodes

Solar panels are basically a sandwich consisting of a metallic substrate that acts like one contact and a transparent substance at the other side with metallic strips to act as the contact to the other side. Inside the sandwich we grow an n-type semiconductor on top of a p-type semiconductor. This forms a very large array of pn-junctions that works the same way as I explained in Section 7.1. Parts of the panels can be damaged or partially obscured by leaves or clouds or a passing squirrel. We can use diodes to bypass damaged solar cells (see Figure 7.12).

Solar panels are connected in series to increase their total voltage and in parallel to get more power. Suppose that the second panel in the first row in Figure 7.12 is defective and for some reason does not work so its resistance is high. Diode D12 becomes forward biased, that is, the contact to the left of D12 has a higher voltage than the contact at the right and therefore D12 shorts and bypasses the damaged panel. All the other diodes are open, thus the current in the upper line goes through the first panel, then up through diode D12 and back down through the third panel. All the diodes on the lower line are open and the current goes through all the three panels. The two diodes at the end of the lines prevent the current from going back if an entire line of panels does not work. This scheme provides power at a reduced voltage rather than no power at all.

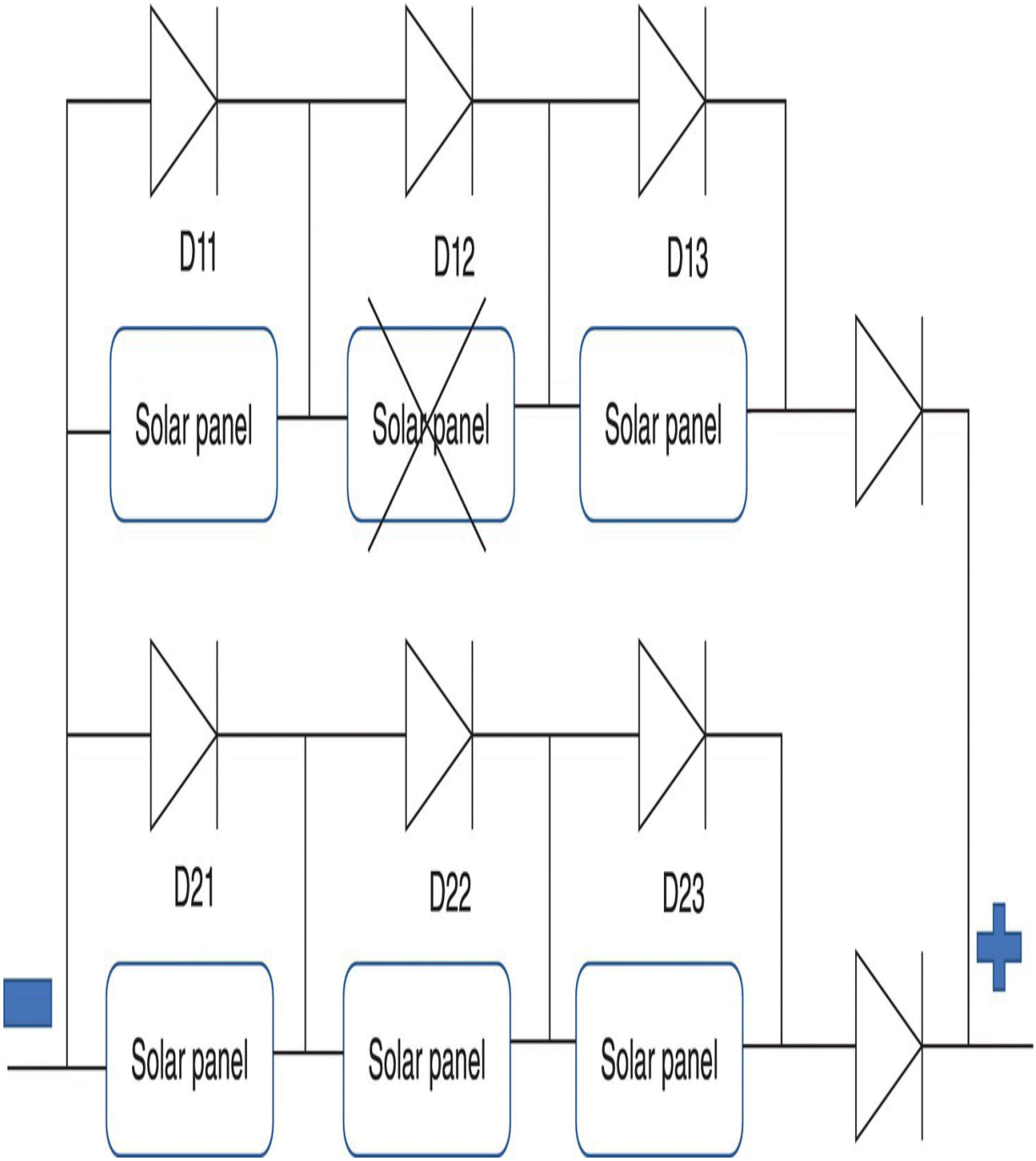


Figure 7.12 Diodes are used to bypass damaged solar cell panels.

7.8 Applications of Schottky Diodes

As I mentioned in the previous chapter, Schottky diodes have the advantage of a very low turn-on voltage and very fast switching speeds, nanoseconds instead of microseconds for standard

semiconductor diodes. This ability to change from a conducting to a nonconducting state is usually called the recovery time. As a result, Schottky diodes are much more efficient than regular diodes when we need higher speeds.

As far as applications are concerned, these are the same as for regular semiconductor diodes. We use Schottky diodes for clamping circuits and for current protection circuits. Their lower turn-on voltage means that there is less power dissipation. You can gather then that, because of their fast speed, Schottky diodes are very useful in all types of logical circuits. I discuss logic circuits in Chapter 11.

7.9 Applications of Zener Diodes

The main advantage of Zener diodes over regular semiconductor diodes is that they can be designed to have a very specific, controlled reversed voltage and can operate in reversed breakdown mode without burning up (as long as we keep the current limited to a certain value). You can buy Zener diodes with *Zener voltages*, that is, the breakdown voltage, in a large range of voltages, from 1 to 300 V, and tolerating currents from a few nanoamps to up to a couple of amperes. (You can peruse the variety of Zener diodes available at a supplier such as http://mauser.com.)

This ability to keep the voltage accurate and constant makes Zener diodes ideal for some of the applications that I have described above. Consider, for example, a voltage clipper (Figure 7.13). This circuit does exactly the same thing as the clipping circuit I showed in Figure 7.8, but I do not need to add any batteries. I just choose a Zener diode with the voltage rating I need and, as I said above, the selection is almost limitless. Quite a simplification. Additionally, we do not want to add batteries inside a microchip.

By this time, you don't need much explanation on how the circuit works. Suppose I want to limit the input voltage of 10 V to a positive swing to 5 V and a negative swing to −2 V. I choose D1 to be a

Zener diode rated at 5 V and D2 to be a Zener diode rated at 2 V. When the positive input voltage cycle is less than 5 V, both diodes are reversed biased and thus not conductive. The output voltage is equal to the input voltage since there is no current, and thus no voltage drop, across the resistor. But when the input voltage reaches or exceeds 5 V, diode D1 is forward biased, and therefore shorted, clamping the output voltage to 5 V. The same is true when the input voltage becomes negative except that now D2 is shorted when the voltage reaches negative 2 V. The resistance I show limits the current so that the diodes do not burn.

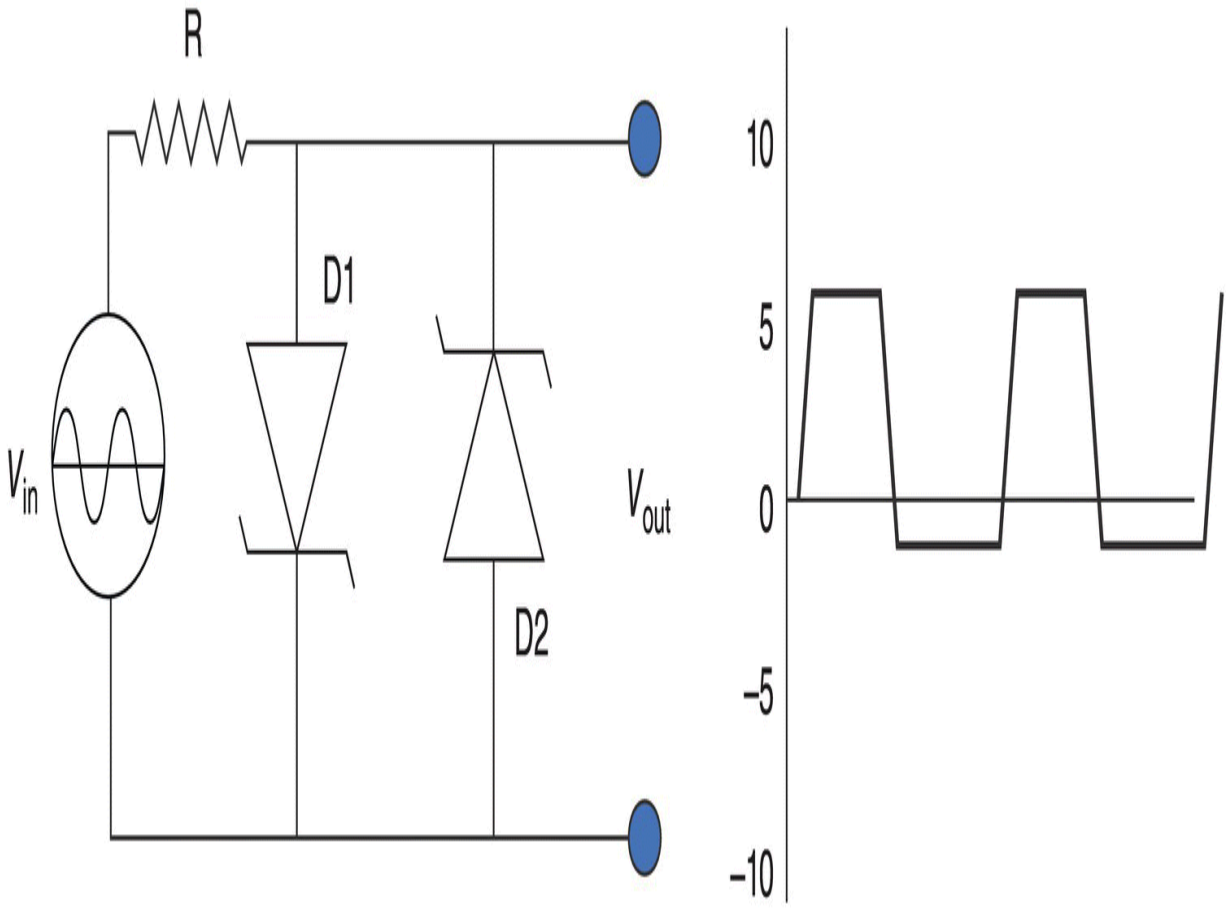


Figure 7.13 A voltage clipper using Zener diodes can clip the voltage depending on the selected voltage rating of the diodes.

7.10 Summary and Conclusions

In this chapter we have not advanced our knowledge of semiconductor device theory, but we have learned some of the very important uses of diodes, from solar cells to rectifiers, and different ways the output voltage can be modified or limited.

So we have seen quite a variety of uses for the simple semiconductor diode. Time to go on to the next and most important semiconductor device, the transistor.

Appendix 7.1 Calculation of the Current Through an RC Circuit

For those interested in looking a little more closely at how a capacitor discharges through a resistor and thus makes the current through the resistor in Figure 7.3 more uniform, more like DC, and the effect of the value of the resistor and capacitor on how well the output voltage will be "constant", I will show you some of the reasoning.

Figure 7.14 is the same as Figure 7.3 but I have replaced the diode by an open switch, so I consider the case when the diode is reversed biased. We know that when the input voltage was positive, the diode was forward biased, on, and charged the capacitor almost instantaneously to the same voltage as the input. As soon as the input voltage starts decreasing, the voltage in the capacitor, due to the charges stored, is larger than the sinusoidal input voltage, which has started going down. This is the situation I show in Figure 7.14. I replace the diode by an open switch, and I have identified the voltage at the capacitor as $v_C(t)$, using a lower case v to emphasize the this is not a constant voltage, but a voltage that changes as a function of time.

At the instant I turn the switch off, $t = 0$, the voltage across the resistor and the capacitor, which must be the same, is the peak voltage of the input voltage, let me call it V_{iMax}. So, I can write:

$$v_C(t) = iR \text{ or } \frac{q(t)}{C} = iR \tag{7.1}$$

since the voltage across the capacitor, $v_C(t)$, is equal to the charge across the capacitor, which is now a function of time, divided by the capacitance. To solve for the current, $i(t)$, as the current is also now a function of time, I divide both sides of the equation by dt, and get

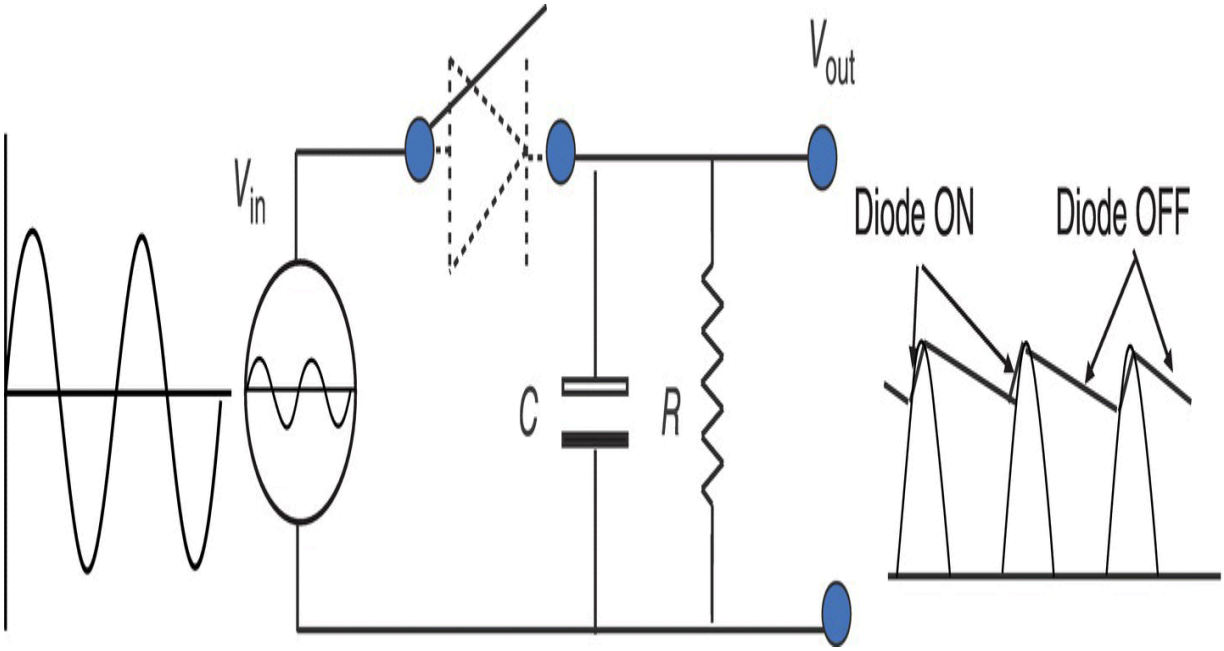


Figure 7.14 An equivalent diode rectifier circuit when the diode is reversed bias.

$$R\frac{di}{dt} == \frac{1}{C}i \quad \text{or} \quad R\frac{di}{i} - \frac{dt}{RC} = 0 \tag{7.2}$$

Many of you will recognize Eq. (7.2) as being a simple differential equation, which can be solved to give

$$i(t) = \frac{V_{i\max}}{R} e^{-t/RC} \tag{7.3}$$

Let's see if Eq. (7.3) makes sense. At $t = 0$, e to the 0 power is 1, therefore the current is the maximum input voltage divided by the

resistance. At $t = \infty$,e to the minus infinity is zero. So, as we expected, the current, and therefore the output voltage in Figure 7.14, starts with a voltage equal to the input voltage and decays to zero. How fast it goes to zero depends on the product of resistance times the capacitance. The higher the product, the faster the current goes down.

Suppose that we want the voltage across the resistor to decrease to no more than 90% of its maximum value. If we go to the table of exponential function e, we see that

$$e^{-0.1} = 0.905 \tag{7.4}$$

The cycle time is 60 Hz or $t = 0.0167$ s. So we want the product of RC to be greater than 0.167, say 0.2, so if the resistance is 1 kΩ, the capacitor needs to be 0.2 mF, which is quite a large capacitor, but if the resistance is 1 MΩ, the capacitor needed is 1000 times smaller, 0.2 μF. These combinations of resistances and capacitances assure us that the output voltage will be less than 90% of the input voltage.

8
Transistors

OBJECTIVES OF THIS CHAPTER

In Chapter 5 I explained the fundamental property of semiconductors devices, the pn-junction. All the other devices that I discuss in the rest of the book are based on different combinations of semiconductor and metallic materials forming pn-junctions. If you have any doubts about the operation and theory of pn-junctions, please review Section 5.1 (and if you feel so inclined, the two appendices in Chapter 5). The key concept is the fact that, in a pn-junction, electrons from the n-type material move to the p-type material because of their density difference, the diffusion current, and the holes move in the opposite direction, resulting in the n-type semiconductor becoming more positive, because it lost electrons, and the p-type more negative for exactly the opposite reason, it has gained electrons. This internal voltage difference results in an opposite current, the drift current, due to the internal electric field, created by the gain and loss of elections, until the two currents, diffusion and drift, cancel each other out.

In this chapter I discuss what happens when we have not two, but three layers of different types of semiconductors and we fabricate the pnp- or npn-transistor, and some modifications such as using metals in conjunction with semiconductors to obtain the complementary metal oxide semiconductor (CMOS) transistors so fundamental in today's electronics. Forbes, in 2014, estimated that we have fabricated 2 913 276 327 576 980 000 000 transistors. That was six years ago as of the time of this writing. How they got an accuracy of 15 significant figures, I do not know.

Let see how a transistor works, what its electrical characteristics are, and how we use it to amplify or control signals. After a fluidics analogy, I explain how a junction transistor and a field-effect transistor work.

8.1 The Concept of the Transistor

A transistor is a device where one current or voltage controls the flow of current in another part of the device. It get its name, “transistor”, from “transforming resistor,” that is, a resistor that I can control its value.

Let me have some fun and consider the fluidic circuit analogy in Figure 8.1. First look at the diagram on the left. There is a channel with a gate separating the left channel, which is filled with water, from the right channel, which is empty. A bypass pipe takes some of the water, but the faucet is closed, so no water flows through this pipe either. The gate is attached to a lever, normally closed by a spring pushing it down.

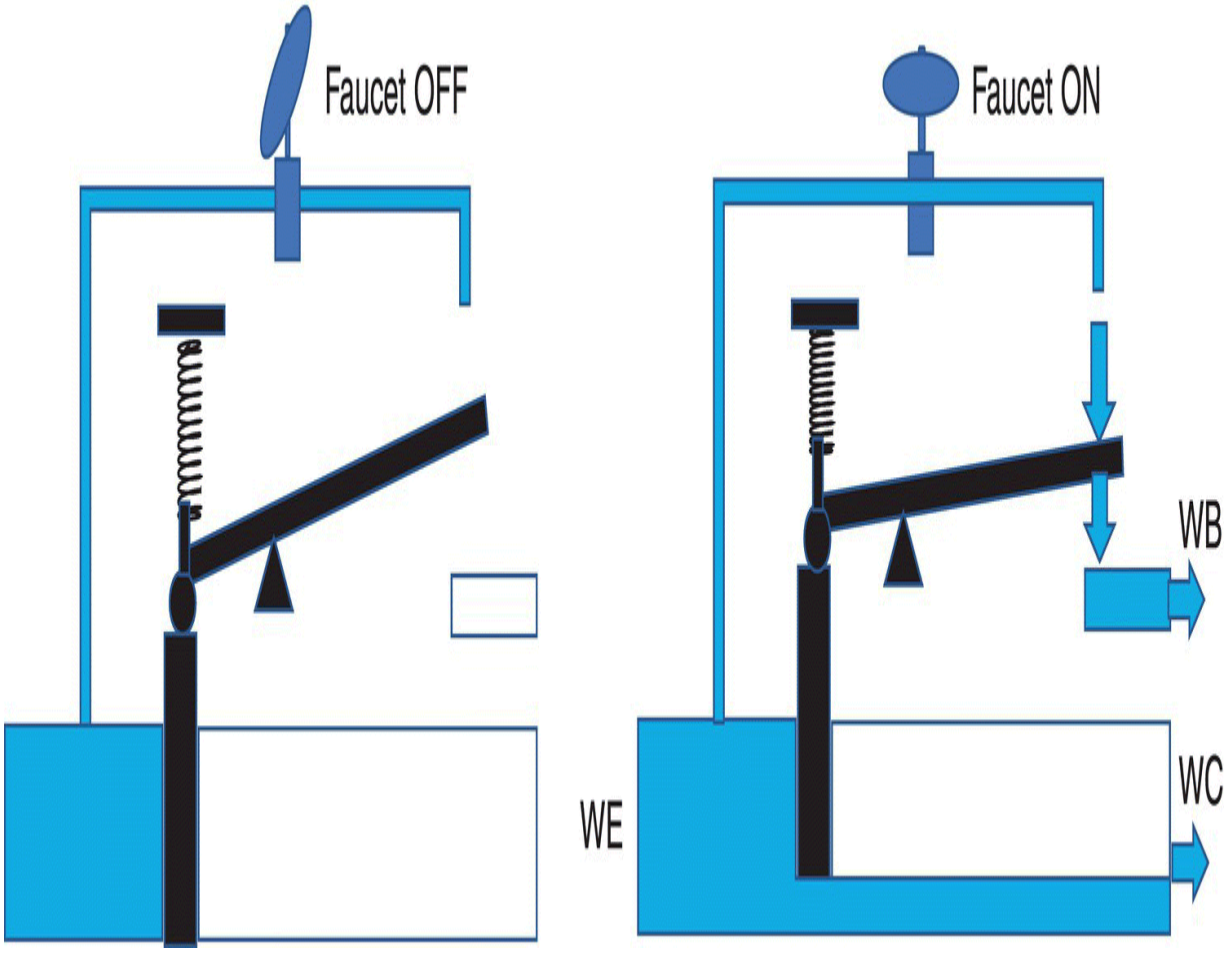


Figure 8.1 A small water flow on the upper pipe controls a much larger flow of water in the main channel.

In the diagram on the right, I open the faucet a little (or a lot) and a small (or large) amount of water runs through the thin upper pipe. I call this flow WB. This flow in the upper pipe is enough to push the lever down, opening the gate and letting some water flow onto the main circuit. I call this flow WC. The amount of water flowing from the left to the right depends on how much water pushes the lever down. The concept here is that a small flow of water in a part of the fluidic circuit can control a much larger flow in the main channel. It is much easier to turn a faucet off or on, even just a little, than to raise the gate in a controlled manner. As long as there is water flowing in the small pipe, the gate remains open and the water in the main channel continues flowing. The more I open the faucet, the more water drops on the lever, the more the gate opens, and the

more water flows to the right channel. A very small amount of the water on the upper pipe controls a large flow of water in the main channel. This is what a transistor does, but with electrons instead of water. A small flow of electrons in one part of the circuit controls a large current in another section of the same device. Also notice that the water coming in from the left must be equal to the sum of the water coming out, so WE = WB + WC (after reading the next section you'll understand why I used this terminology for the water flow).

There are two types of transistors, the bipolar junction transistor (BJT) and the field-effect transistor (FET). There are also two types of FETs. One, which we will call the junction field-effect transistor (JFET), uses only n- and p-type semiconductor materials. The other type is the metal-oxide semiconductor FET (MOSFET), which replaces one type of semiconductor material for a metal. I explain all three types in this chapter.

8.2 The Bipolar Junction Transistor

This semiconductor transistor consists of a very thin p-type semiconductor sandwiched between two n-type semiconductor materials. We call this structure an npn-transistor. As you would expect, we can also have a very thin n-type material sandwiched between two p-type semiconductors, which, obviously, we call it a pnp-transistor. Both work the same way.

You should recognize that we have two pn-junctions, back to back, with a very thin p-region in the middle. At room temperature and without any external bias, the electrons from the n-type semiconductors on both sides of the base diffuse into the p-region and holes from the p-region move to both n-type materials, generating two transition regions, exactly as happens with a single pn-junction. The n-type regions that have lost electrons to the p-type base become more positive and the p-region that has gained electrons from both sides becomes more negative. I show this electrical potential at the bottom of Figure 8.2. As in the pn-junction, the equilibrium condition occurs when the diffusion currents due to

the difference in electron and hole densities exactly cancels the drift currents due to the electrostatic potentials created at the junctions. We call the center region the *base*, one of the n-type materials the *emitter*, and the other the *collector*. In principle the collector and emitter can be reversed. You will see why we use this terminology in a minute.

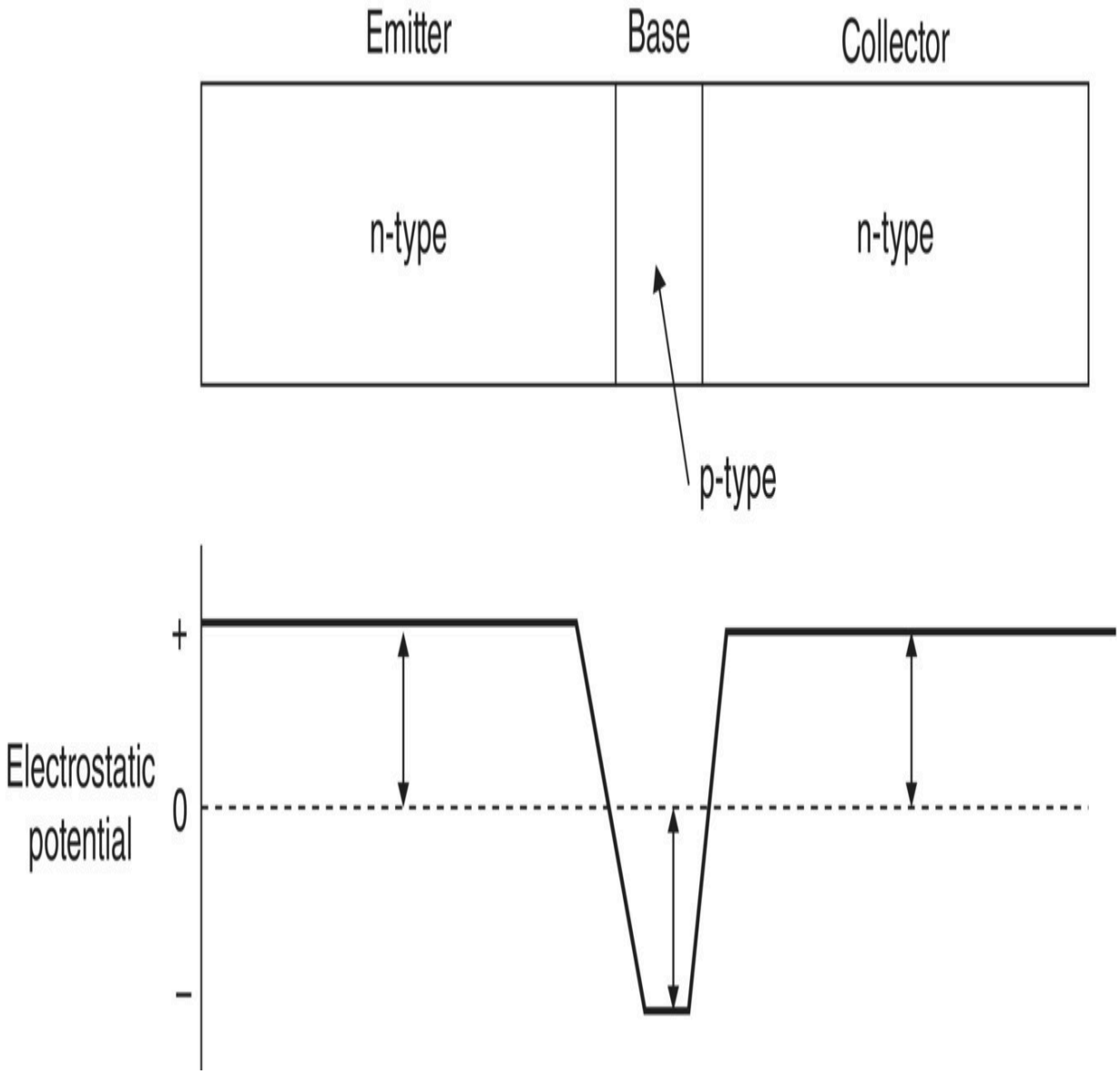


Figure 8.2 The structure of an npn-transistor consists of a narrow p-type semiconductor sandwiched between two n-type semiconductors. This structure creates, like with the pn-junction, an internal electrical potential in both junctions generated by the diffusion of electrons from both sides.

Now suppose we forward bias the emitter with respect to the base with a voltage V_E, E because it is the voltage applied to the emitter. At the same time, we reverse bias the base with respect to the collector with a voltage V_C, C for collector (Figure 8.3).

As I show on the left of Figure 8.3, the electrostatic potential between the emitter and the base decreases by voltage V_E, forward bias. Therefore, a large electron diffusion current flows from the emitter to the base. If the other n-type material, the collector, was not there, the current would flow through the base to the positive side of the battery V_E, the same as happens in a forward biased diode. But now look what happens at the collector side. The pn-junction is reversed biased. There is a large added positive voltage provided by battery V_C. So, if there was no emitter, no electrons would come from the base because it is a p-type semiconductor and no electrons would go to the base because they are attracted to positive side of battery V_C, and as in the reversed biased diode no current would flow.

But when the base is very thin, the electrons coming from the emitter to the base by diffusion enter a region where they encounter a very strong positive electric field attracting them to the right, that is, to the collector. Most of the electrons don't have time to escape down the base to the positive side of V_E before they are attracted by the large positive potential just across the base and move directly to the collector and to the positive terminal of the battery, V_C. Yes, a small number of electrons flow through the base to the positive terminal of battery V_E, but the thinner we make the base, the more electrons are swept by the strong positive potential at the base–collector transition region.

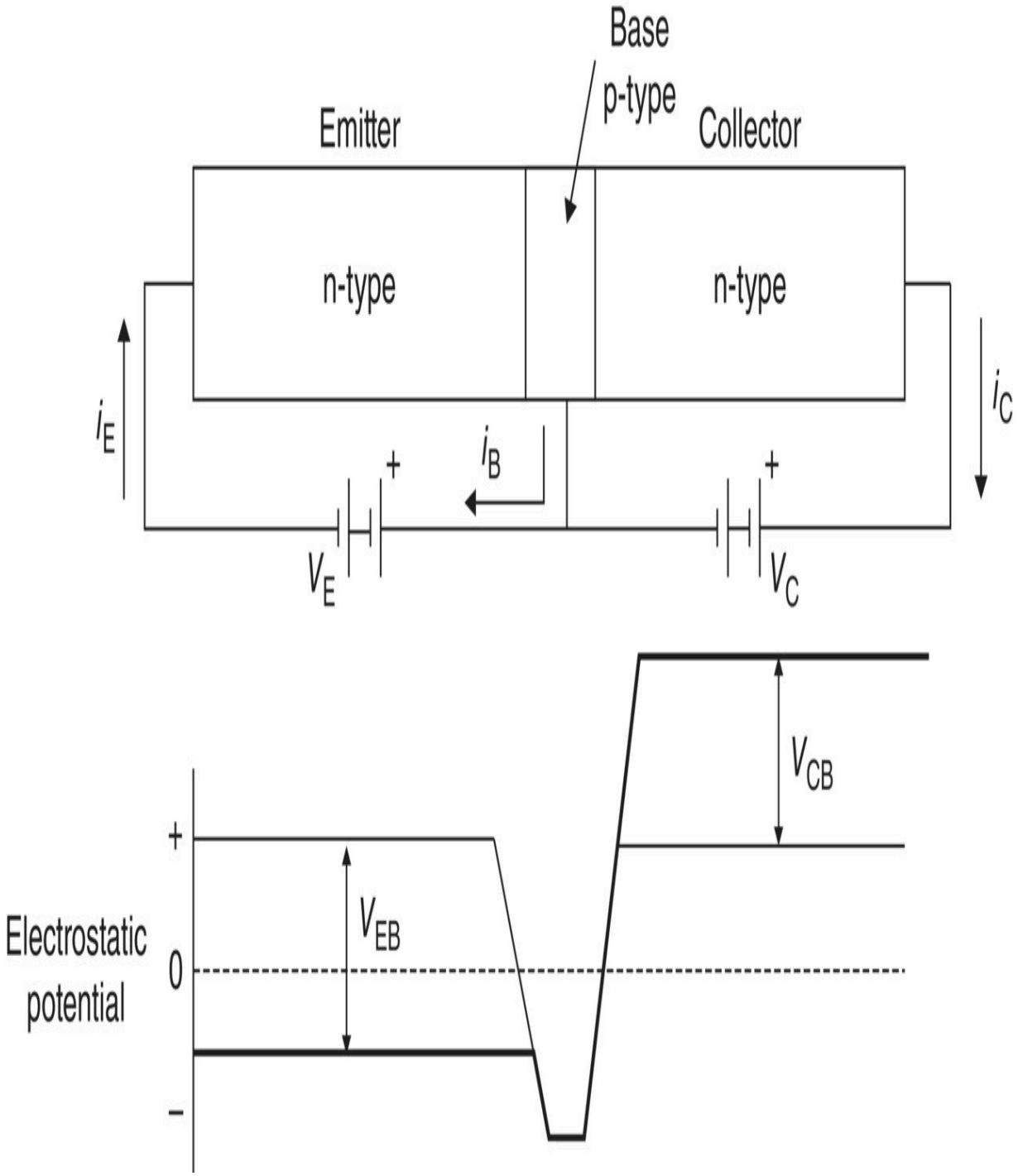


Figure 8.3 When we apply external voltages to an npn-transistor the internal electrical potential changes, making the emitter more negative (forward bias) and the collector more positive (reversed bias).

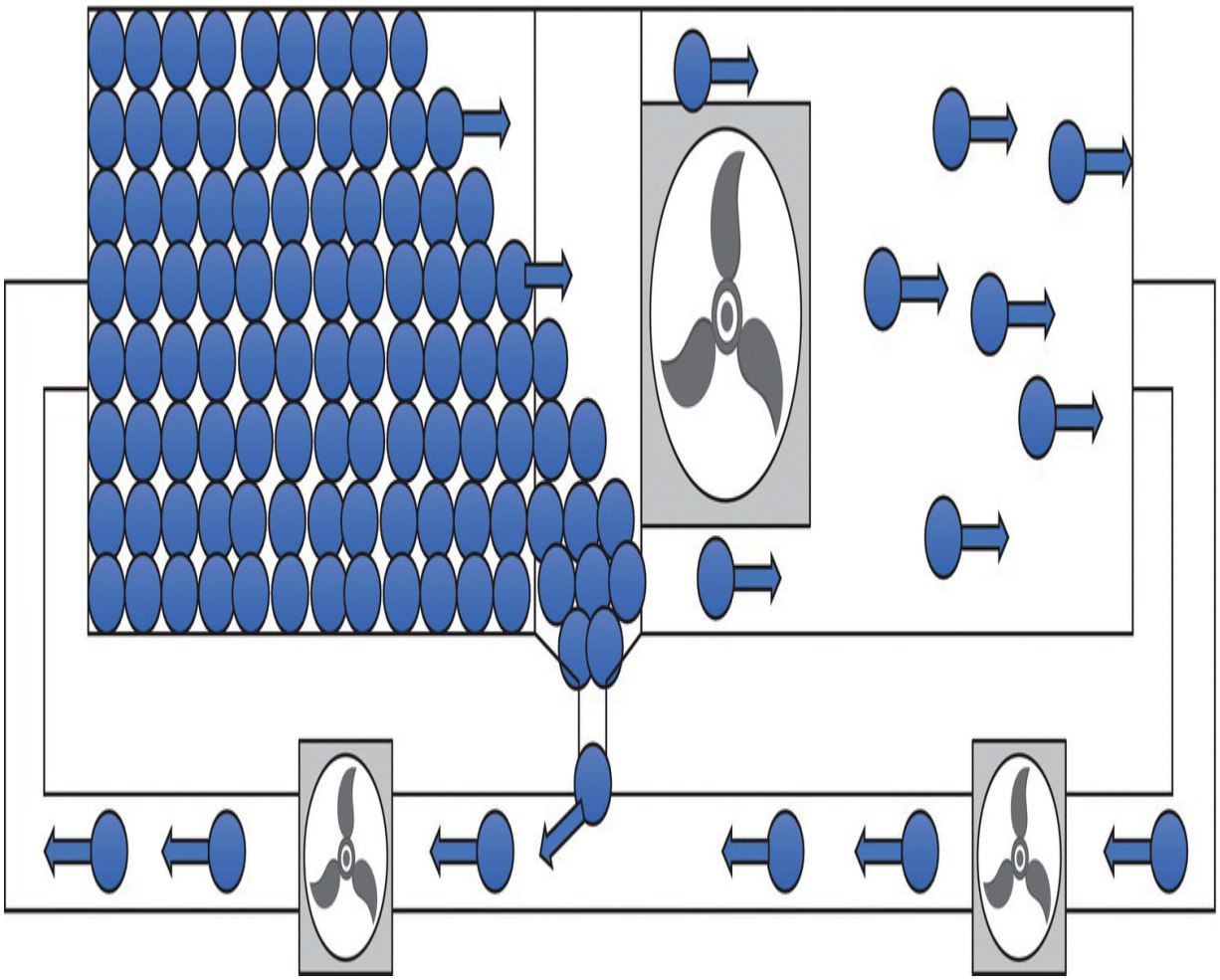


Figure 8.4 Some balls fall from a box full of ping-pong balls (electrons) on the left to the empty center region. Some of the balls fall through the bottom of the center region but the majority are swept to the right by a large fan (the electrical potential).

I show this situation in Figure 8.4 using another analogy. Suppose I have a box full of ping-pong balls on the left and a thin empty box in the middle. I call the box on the left the emitter box, the one in the middle the base box, and the one on the right the collector box. The balls start falling, diffusing, into the thin box. Some of the ping-pong balls will fall through the middle hole. But now suppose I turn on a big fan, equivalent to the collector voltage in Figure 8.3, on the right of the base. The ping-pong balls that moved to the base from the emitter will rush to the right and will not give time to many balls to fall through the opening at the base.

This is a crude analogy, but it is very similar to what is happening in the transistor. When we bias the transistor, forward biased the emitter junction, and reversed bias the collector junction, the electrons rush from the emitter to the base due to diffusion, but before they have any time to recombine with holes or escape through the base contact, the electrons are swept by the high electrical field between the base and the collector.

As I said with the fluidics example, the emitter current must be equal to the sum of the collector current plus the base current, so

$$I_e = I_c + I_b \tag{8.1}$$

Let's now define a constant, *a*, that tells me what percentage of the emitter current goes to the collector. Then

$$I_c = \alpha I_e \tag{8.2}$$

where *a* is always less than one. Therefore, combining Eqs. (8.1) and (8.2) the base current must be

$$I_b = (1-\alpha) I_e \tag{8.3}$$

For example, if 95% of the emitter current crosses to the collector, then the constant $a = 0.95$, and by necessity the base current is only 5% (1 – 0.95) of the emitter current. These are the key relations of the currents in a transistor. By the way, now you can see why I call the n-type semiconductor on the left of Figure 8.3 the emitter and the one on the right the collector. I show graphically these three currents in Figure 8.5.

Now, if $I_b = (1 - a)I_e$, then

$$I_e = \frac{I_b}{1-\alpha} \tag{8.4}$$

and combining Eqs. (8.2) and (8.4), the collector current is

$$I_c = \alpha I_e = \frac{\alpha I_e}{1-\alpha} = \beta I_e \tag{8.5}$$

where now we have defined a new constant, β,

$$\beta = \frac{\alpha}{1-\alpha} \tag{8.6}$$

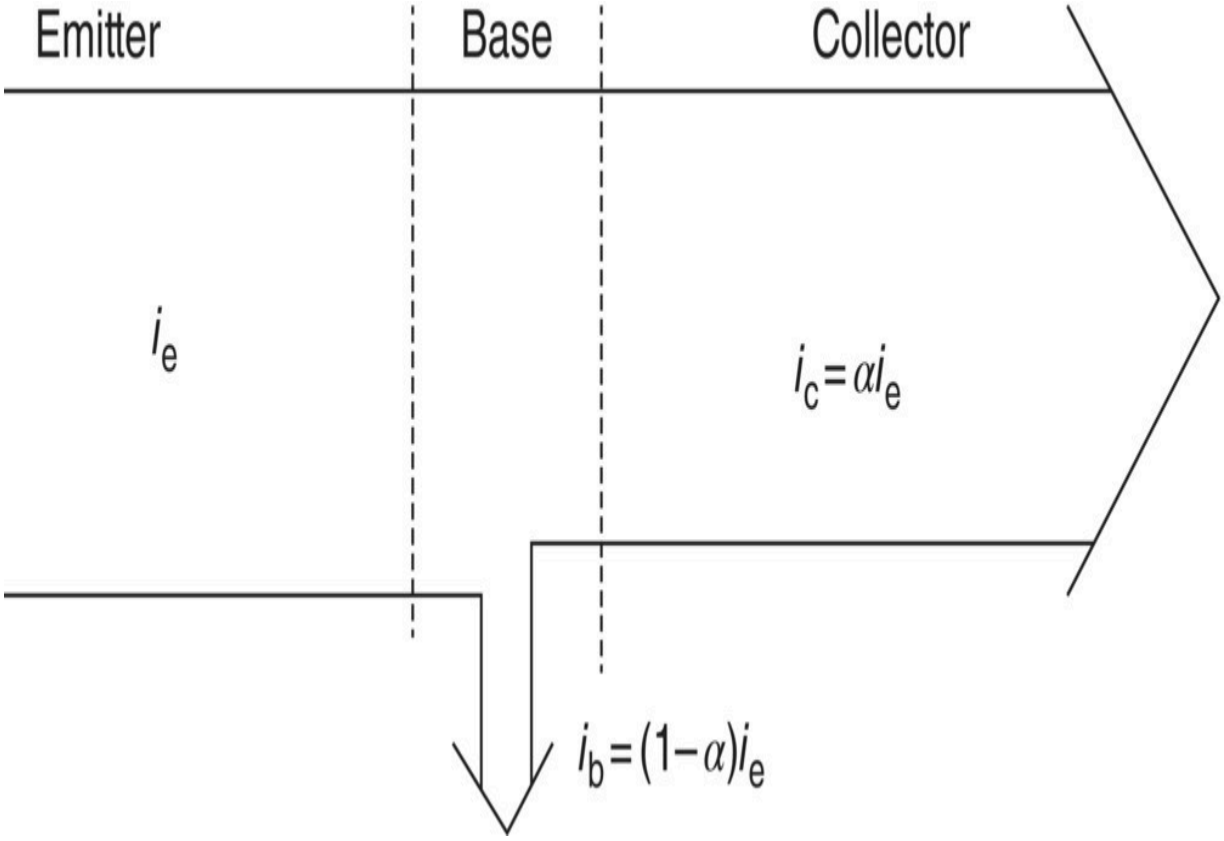


Figure 8.5 The collector current, I_c is proportional to the emitter current, I_e, by a percentage factor *a*. The emitter current must be equal to the sum of the base and collector currents.

In the case where *a* is 0.95, this means that

$$\beta = \frac{0.95}{1-0.95} = \frac{0.95}{0.05} = 19 \tag{8.7}$$

All these relationships tell us something that is quite obvious just by looking at Figure 8.5. The collector current, I_c, is considerably larger

than the base current, in our case it is 19 times larger. (You see what I told you in the introduction. Math is trivial. We just make it look complicated to keep our jobs! You do not need much math to know that 95 is 19 times larger than 5.)

Now if we make the base thinner so that only 2% of the electrons flow through the base, then β would be 49, that is, the collector current is 49 times larger than the current in the base. We call β the transistor gain.

Look at Figure 8.6. What happens if I add a small sinusoidal current to the base of the transistor? Remember the collector current, in our case, is 19 times larger than the base current, so if the base current changes sinusoidally by one unit, the current in the collector has to change by 19 units. If the sinusoidal signal is very faint music, for example, the sinusoidal music wave coming from the collector will be 19 times louder. Quite a trick.

Note that the sinusoidal signal connected to the base must be small compared to the direct current biases, V_E and V_C. We do not want to change the biases between emitter and base, and base and collector so much that the emitter junction becomes reversed biased or the collector junction forward biased or even close to it. That would completely stop the transistor action. If you throw a small pebble into a pool, you'll see water waves moving radially, neatly traveling to the edge of the pool, but if you throw a large boulder, the water will move chaotically and splash all over the pool. To observe the ripples in the water, the disturbance has to be small so that the basic conditions of the water, that is, the desire of the water to be in its lowest possible equilibrium state, is not disrupted.

There is an electronic symbol for the transistor (Figure 8.7). I show the emitter with an arrow indicating if the positive current goes into, pnp, or out of, npn, the transistor's emitter. The base is the contact in the middle and the collector just accepts the charges that are coming from the emitter. Note again that the direction of the current I show in the emitter is the direction of the positive charges.

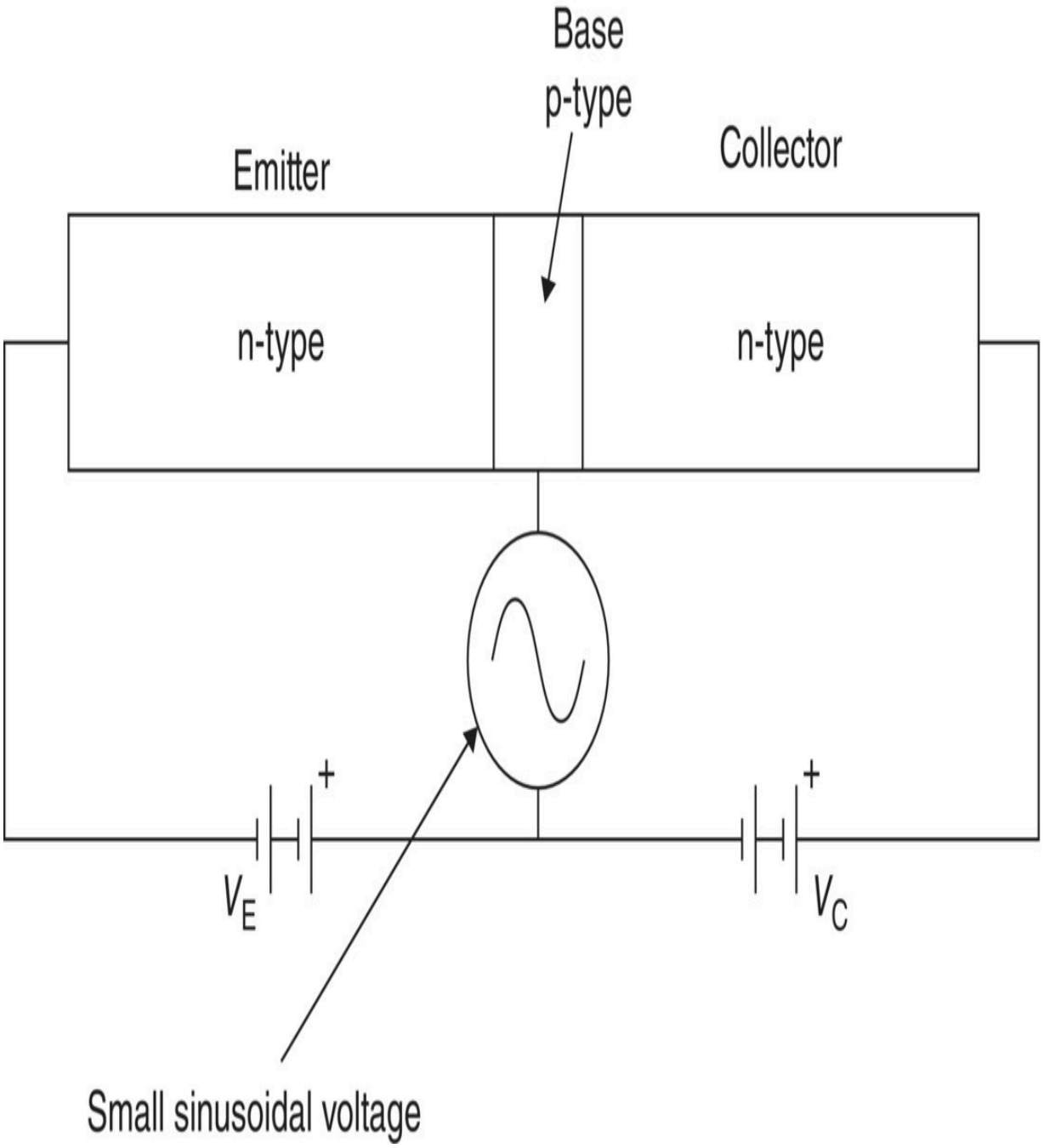


Figure 8.6 Adding a sinusoidal signal to the base of a transistor properly biased with direct current provided by batteries results in a much larger change in the collector current.

Figure 8.8 shows the characteristic curves of a transistor. The characteristic curves graphically show us the relationship between voltages and currents in a transistor. Figure 8.8 shows the collector

current as a function of the applied voltage between the emitter and the collector for different base currents.

The first thing I want you to notice is that even with zero base current (the lowest curve in Figure 8.8) I_{ceo} and $I_b = 0$, the collector increases linearly from zero to a higher value as the collector to emitter voltage increases. The characteristic curves show the collector current, I_{ceo}, going from 0 mA at 0 V to about 0.5 mA at 6 V. We call this the leakage current, I_{CE0}, and it has nothing to do with the transistor effect. The leakage current is due to the fact that the collector to base junction is reversed biased and the base region is thin so that electrons can cross it as the voltage increases. If we did not have this leakage current, all the other curves for I_b from 20 to 100 µA would be horizontal, that is, for a given base current, the collector current is constant and does not change with the increase in voltage. The only reason that all the curves have a slope is because the additional leakage current I_{CE0} is added to the total collector current.

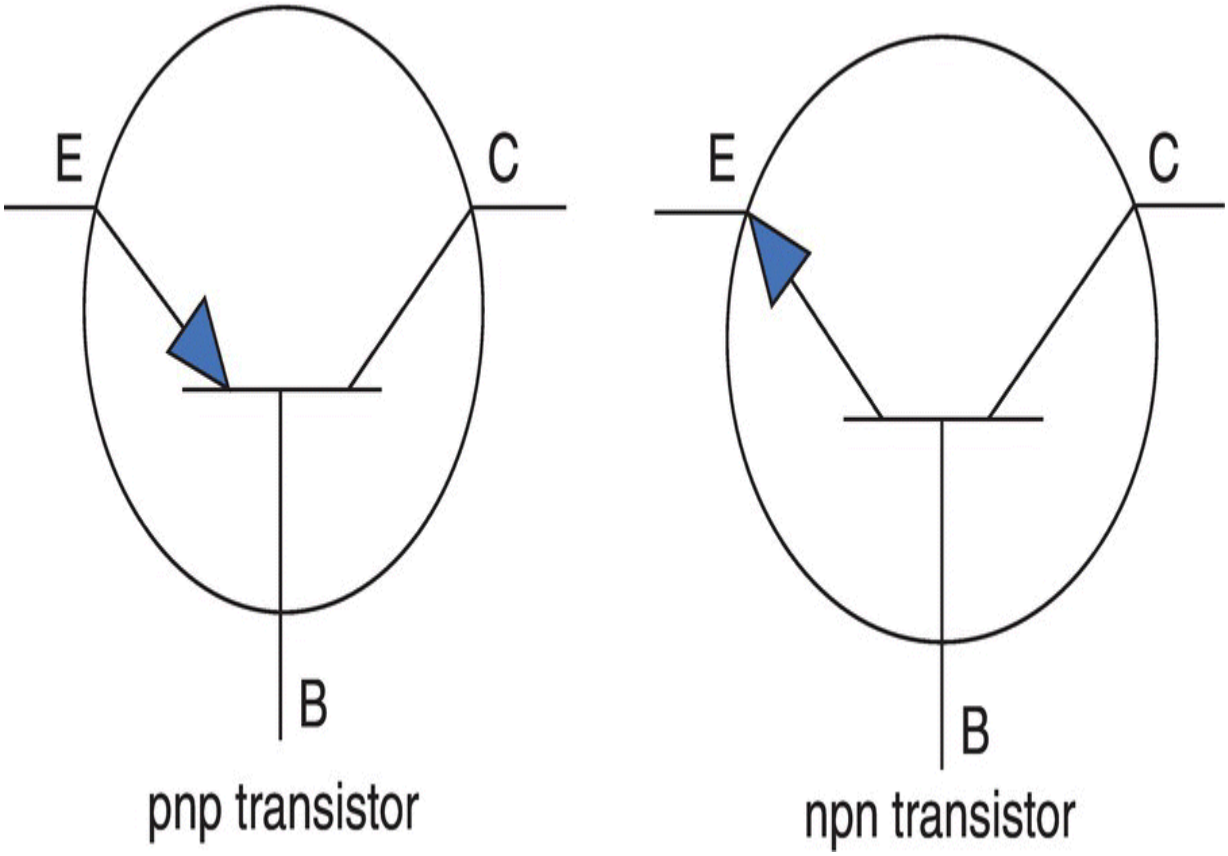


Figure 8.7 Symbols for pnp- and npn-transistors. The arrows show the direction of the positive current in the emitter.

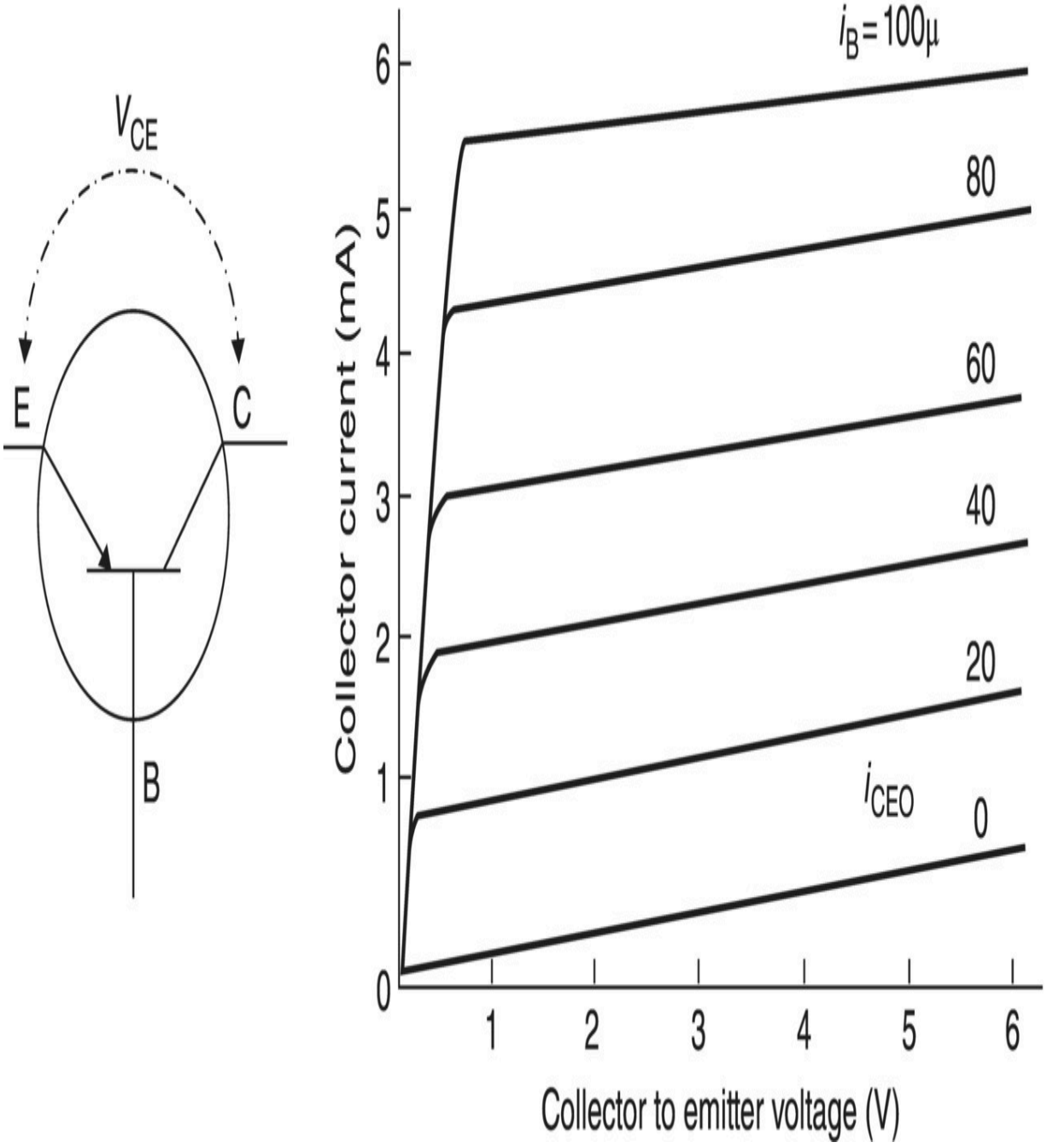


Figure 8.8 Transistor performance is graphically given by the collector current versus the collector to emitter voltage as a function of the base current. The slope of the lines is due to the leakage current, I_{CEO}.

If I keep increasing V_{CB} (see Figure 8.3), at some point the slope of the transition region between the base and the collector will reach

the emitter, connecting both n-type materials, and the current will increase drastically. We call this the breakdown condition.

One important and final point is that both the current I_{CE0} and the transistor gain, β, are not stable. They change with temperature and voltage. This change is quite important if we want a very accurate gain. Have this in mind when I show how to bias a transistor in Chapter 9.

8.3 The Junction Field-effect Transistor

Another transistor device is the JFET, sometimes called the Schottky transistor. It works quite differently from the standard transistor, but its operation depends on two pn-junctions. I show a graphical simplistic representation of a JFET in Figure 8.9.

The JFET is like an Oreo cookie, where the sweet cream filling, the n-type semiconductor, is sandwiched between two chocolate biscuits, the two p-type semiconductors. That's it. To bias it properly, the left side of the n-type semiconductor is grounded and the right side has a positive voltage, V_D. The two p-type semiconductors are connected together and right now they are grounded, that is, voltage $V_G = 0$. This structure results in two pn-junctions, one at the top and one at the bottom of the channel. The difference is that now the current does not go through the junctions, as in the BJT. The current goes only between the junctions through the n-type material from right to left. This transistor is called a *unipolar transistor* because only electrons in the n-type material are moving (or only holes in a p-type channel). The regular transistor, BJT, is called a *bipolar transistor* because both electrons and holes, moving in opposite directions, contribute to the currents. The region in the middle is called the *channel*, the left end of the n-type material is called the *source*, and the right end of the n-type material is called the *drain*. We call the two p-type semiconductors the *gates*. The drain and source could be reversed as long as I bias them differently.

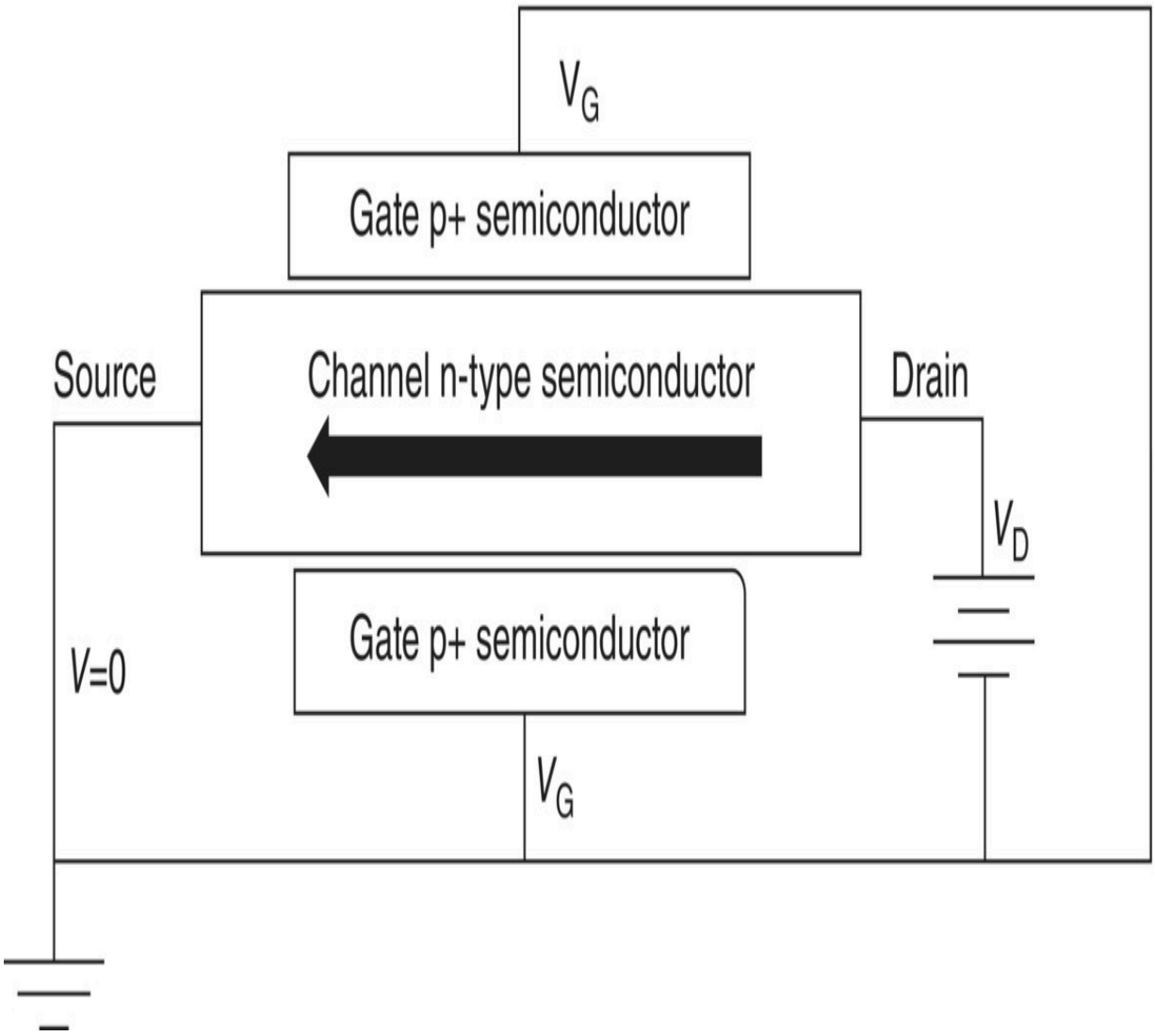


Figure 8.9 The structure of an n-type JFET consists of one type of semiconductor, n, sandwiched between two semiconductors of the opposite type, p.

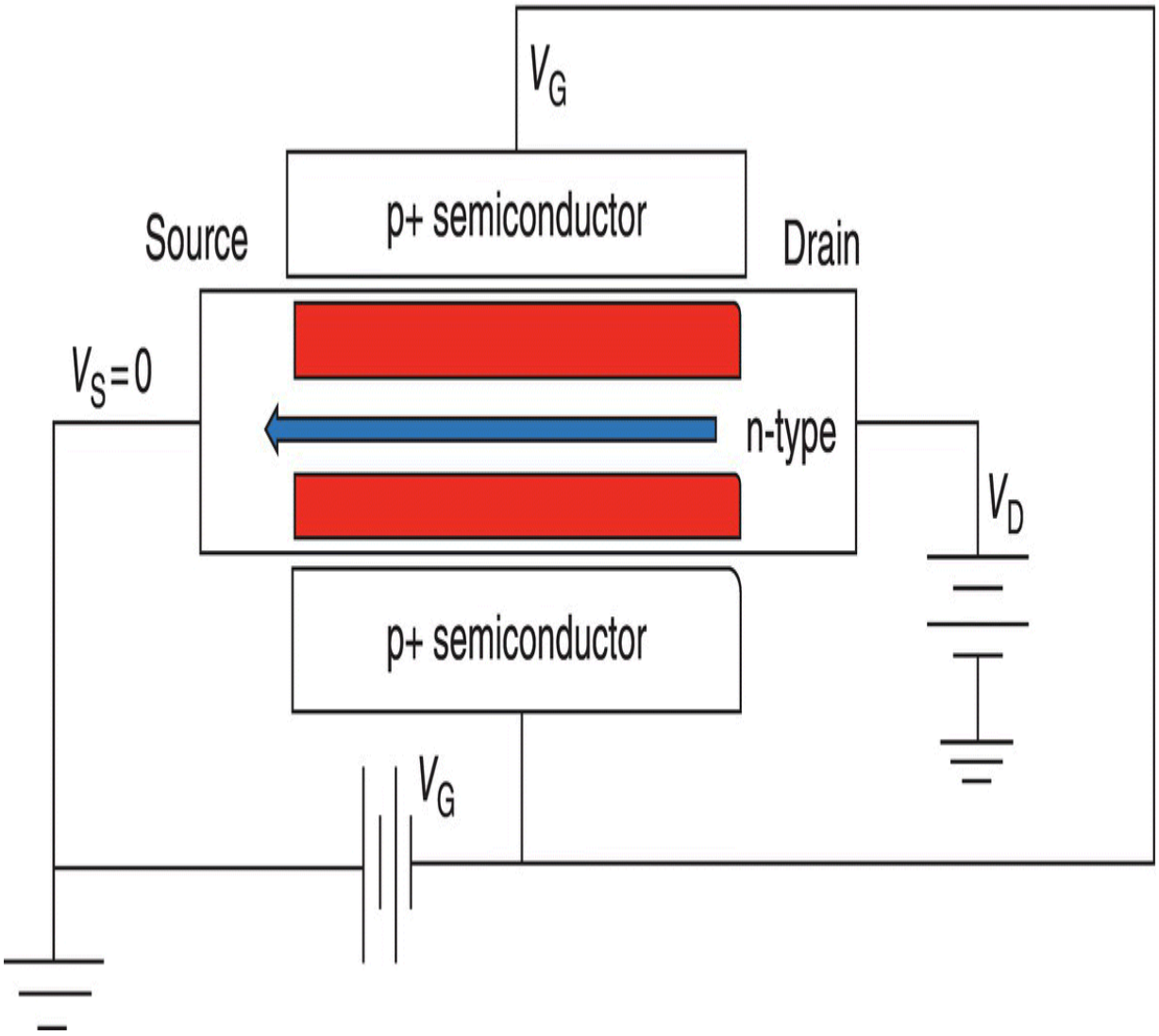


Figure 8.10 A JFET with a positive voltage at the gates creates two depletion regions that makes the conductive part of the base thinner.

Let's see how it works. In Figure 8.9, the gates are grounded so the gates have the same voltage as the source. Suppose now that I apply a negative voltage at the bases, as I show in Figure 8.10.

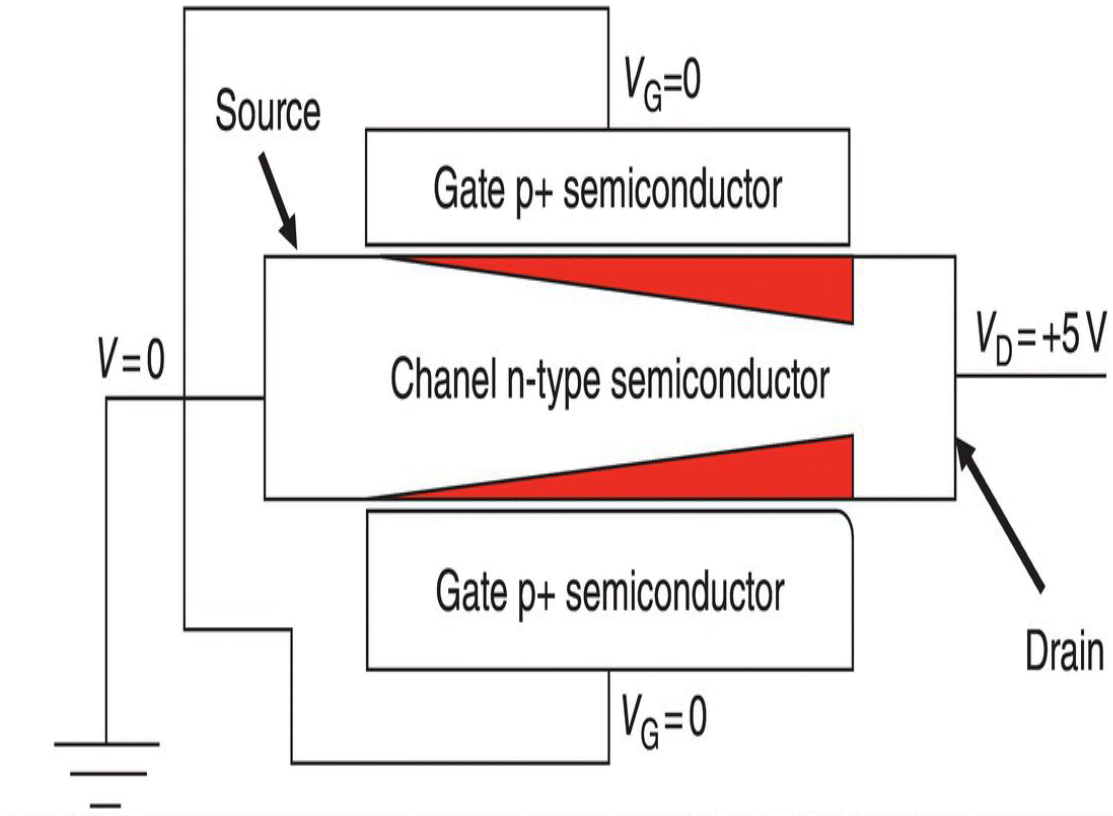
Now I have two reversed biased pn-junctions that repel electrons, thus creating two transition regions, one at the top and one at the bottom. At the transition region, also called the depletion region (Section 5.1), there are no free charges. The channel is now thinner, resulting in a higher resistance than it had when the gates were grounded (Figure 5.9), and if I do not change the drain voltage, V_D,

the current decreases. So, I have a device where I can change its resistance by applying a voltage at the gates.

As you might expect, the depletion region in the middle of the channel is more complex than what I show in Figure 8.10. The problem is that the voltage inside the channel is not uniform. If I set the gate voltage to zero, at one end, the source at the left, the voltage between the source and the gates is zero, and at the other end, the drain at the right, the voltage between gates and drain is V_D. So in between the two ends the voltage across the channel and the gates has to change from 0 to V_D in some continuous and uniform way. We would expect the voltage in the middle of the channel to be about $V_D/2$. How does this affect the depletion region?

Assume first that the voltage at the gate is zero (upper drawing in Figure 8.11), that is, the two gates and the drain are grounded, and we apply a positive voltage at the drain, let us say $V_D = 5\,V$. I show this case at the top of Figure 8.11.

Take a look first at the upper figure. At the left end of the channel, the source, there is no voltage difference between the p-gate and the n-type channel. They are both grounded and the voltage between them is zero. Therefore, the depletion region on the left of the device, is very small. As we go to the right, the drain side, there is a 5 V reversed bias voltage between the drain and the gates ($V_G = 0\,V$ and $V_D = 5\,V$). The depletion region increases and the channel gets narrower. In the middle of the channel where the voltage is approximately 2.5 V between the channel and the gates, the depletion region is thinner than at the drain side but thicker than at the source side. Therefore, you can intuitively see that the transition region goes smoothly from a small value at the source to a larger value on the right.


V_G=0
Source
Gate p+ semiconductor
V_D=+5 V
V=0
Chanel n-type semiconductor
Gate p+ semiconductor
Drain
V_G=0

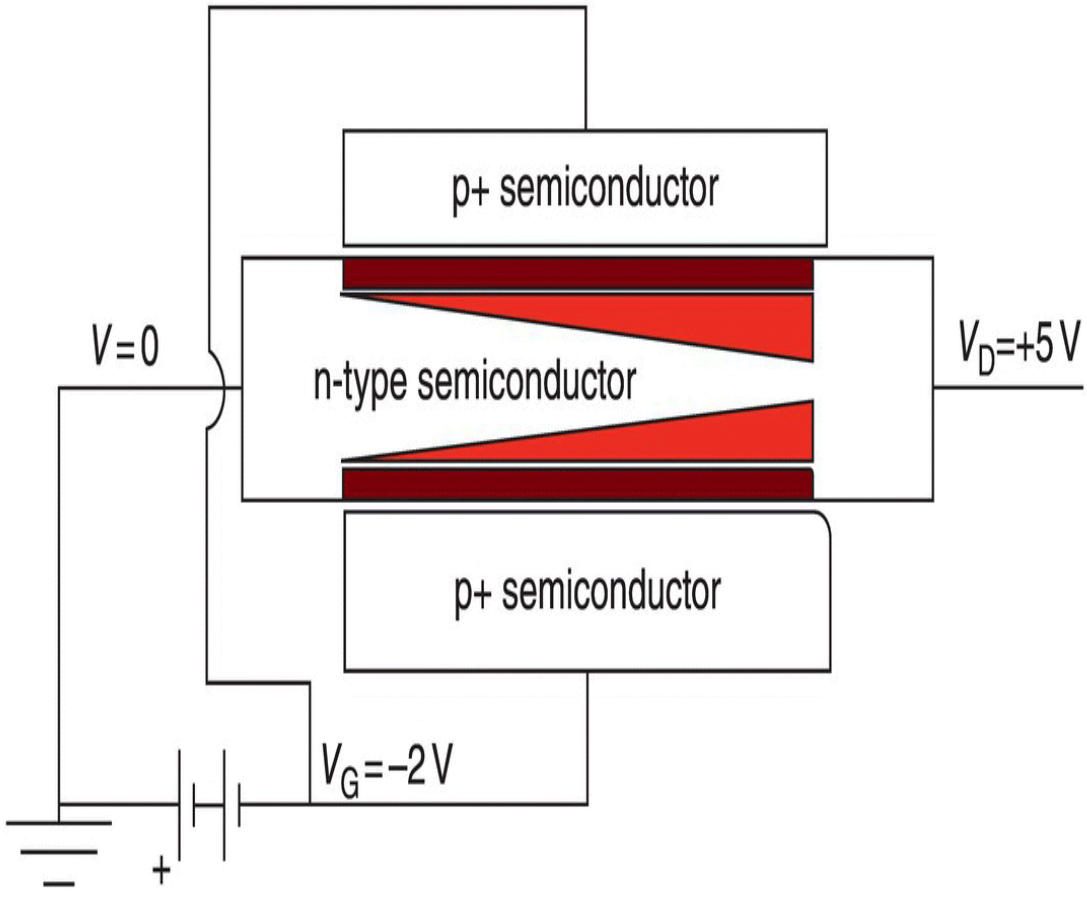

p+ semiconductor
V=0
V_D=+5 V
n-type semiconductor
p+ semiconductor
V_G=−2 V
+

Figure 8.11 The voltage between the drain and the gate is different to that between the source and the gate, therefore the transition region is larger at the drain side and zero at the source side. In the lower figure the transition region has increased uniformly by the 2 volts applied to the gate making the channel narrower.

But now here is the clever bit: if I increase the voltage at the base by −2 V (lower drawing in Figure 8.11), the depletion region increases throughout the gate by a uniform amount, and the path for the electrons narrows still more. If I add a sinusoidal input voltage to the gates, that voltage modulates the active thickness of the channel, the resistance of the channel, and therefore also the current through the JFET. Voltage changes at the gate V_G control and change the dimensions of the path through where the electrons flow, or better, the resistance of the channel changes and the current, for a given drain voltage V_D, also changes. The gate voltage is controlling the current through channel. That is what transistor action does.

The fundamental differences between the BJT and the JFET are as follows:

In the BJT, the current in the base controls the current in the collector. In the JFET the voltage (not the current) at the gates controls the current through the channel.

In the BJT, the current goes *through* the two transition regions. In the JFET, the current goes *between* the two transition regions.

In the BJT, the current is composed of electrons and holes, which is why it is called *bipolar*. The JFET is *unipolar* because only one type of charge, electrons or holes, moves through the channel.

In the BJT there is current though the base. In the JFET there is no current through the gates (just the leakage current).

I have just described what is called the linear region of the JFET operation, linear because as I increase the gate voltage, I increase the resistance of the channel and thus decrease the current. Now

what happens if I keep a constant gate voltage and increase the drain voltage? At some point the depletions at the drain side will try to touch each other. We call this the *pinch-off voltage*. Notice that I use "try to touch" not "touch each other." Let me call this drain voltage the saturation voltage, V_{Dsat}. It looks like at pinch-off no current should flow; the channel is closed. Not so fast. If the pinch-off would shut the current completely, there would be no voltage along the channel, the reversed bias voltage would disappear, the transition region would collapse, and thus the channel will open up. You can see that at pinch-off voltage there has to be a very narrow path so that the transition region does not collapse. But now, no matter how much farther I increase the drain voltage, the voltage between the source and the pinch-off region cannot change, thus the current through the JFET remains constant. Figure 8.12 shows these characteristics.

For a given gate voltage, as we increase the drain voltage (with respect to the source, which I have assumed it is grounded, zero) the current increases. This is the linear or ohmic region (the left-hand region in Figure 8.12), that is, as the voltage increases, the current also increases almost linearly, following Ohm's law. When we reach the pinch-off voltage the current is constant as the drain voltage increases (the middle section) but at some point the voltage is high enough that the junction breaks down (the right-hand region in Figure 8.12).

Let me show another simplified drawing that may help you visualize what is happening with the pinch-off as the drain voltage increases, Figure 8.13.

As an example, I am putting numbers on the contacts to make it, hopefully, clearer. I ask you to go back and forth between Figures 8.12 and 8.13. In all the diagrams, A to D, in Figure 8.13, $V_S = 0\,V$ and $V_G = -3\,V$. (I have highlighted in Figure 8.12 the curve corresponding to $V_G = -3\,V$). The only voltage I change is the drain voltage, V_D.

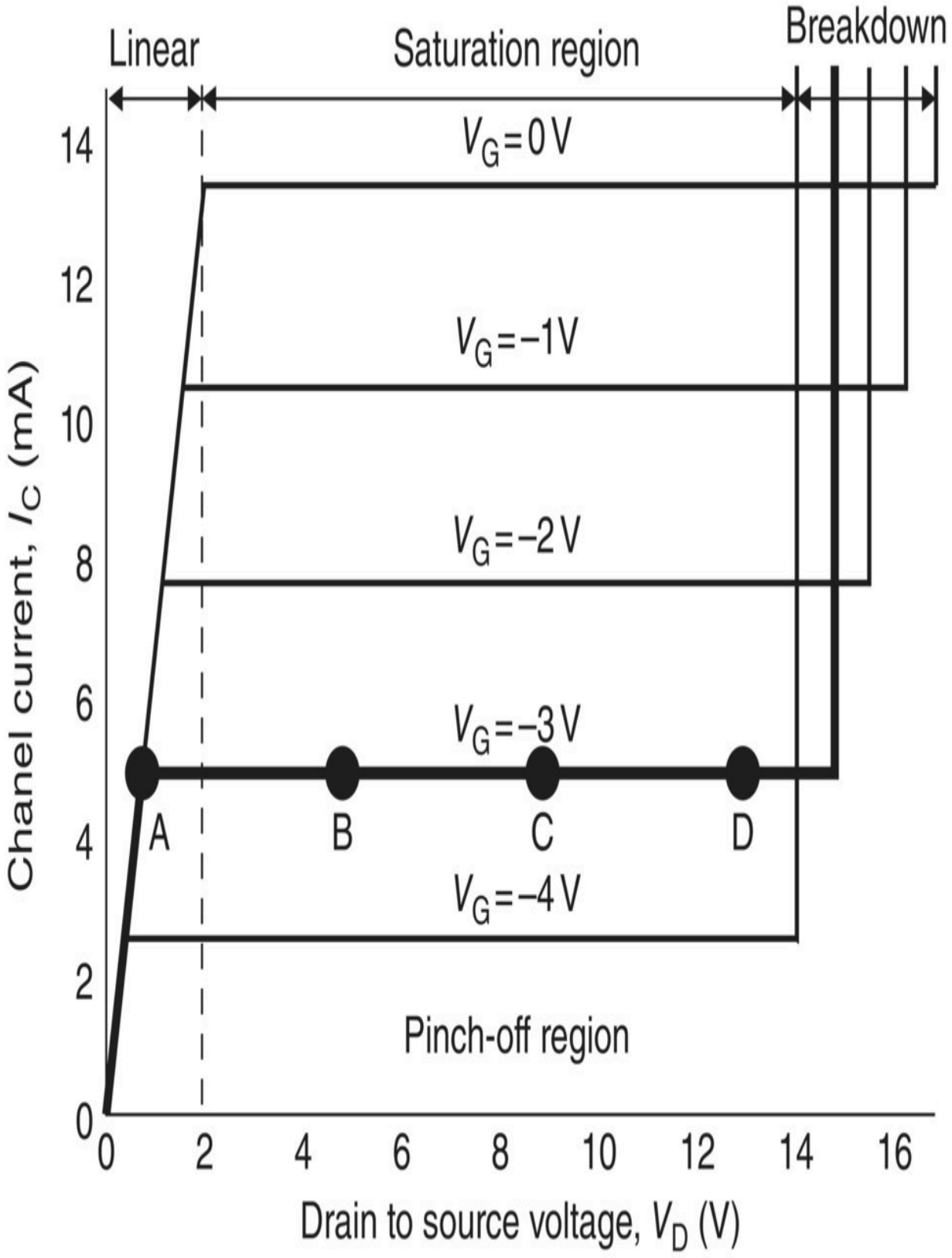

Linear
Saturation region
Breakdown
$V_G = 0\,V$
$V_G = -1\,V$
$V_G = -2\,V$
$V_G = -3\,V$
$V_G = -4\,V$
A
B
C
D
Pinch-off region
Chanel current, I_C (mA)
Drain to source voltage, V_D (V)
0
2
4
6
8
10
12
14
16

Figure 8.12 The idealized characteristics of a pnp JFET show three distinct regions: the linear region where the current increases with drain voltage, the saturation region where the current is independent of the drain voltage, and the breakdown region.

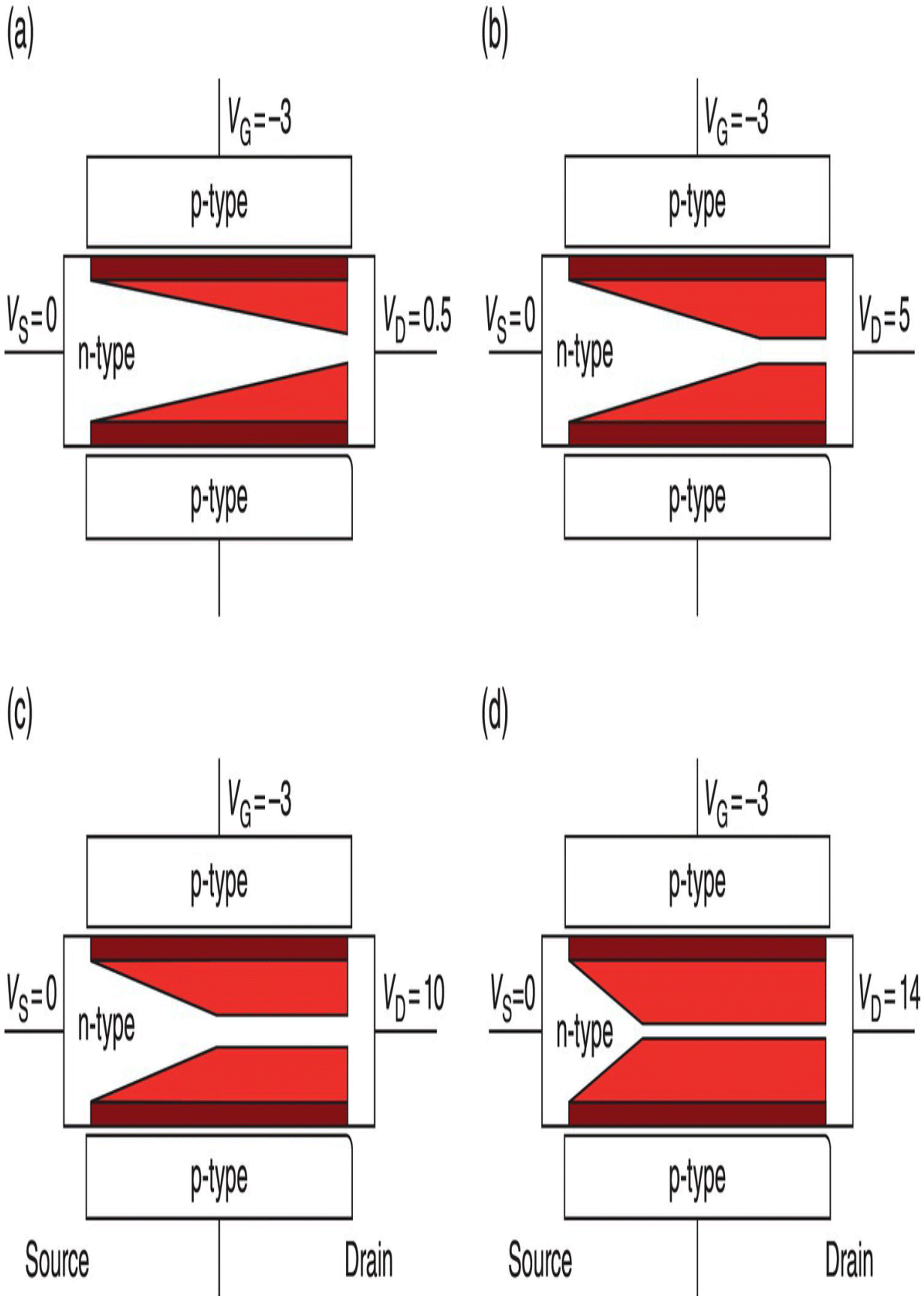

(a)
V_G=−3
p-type
V_S=0
n-type
V_D=0.5
p-type
(b)
V_G=−3
p-type
V_S=0
n-type
V_D=5
p-type
(c)
V_G=−3
p-type
V_S=0
n-type
V_D=10
p-type
Source
Drain
(d)
V_G=−3
p-type
V_S=0
n-type
V_D=14
p-type
Source
Drain

Figure 8.13 The pinch-off voltage grows and moves closer to the source as the drain voltage increases from 0 to 14 V.

When the drain voltage is 0.5 V (Figure 8.13A), it is just sufficient to generate a pinch-off at the drain. As I increase the drain voltage to 5 V (Figure 8.13B), the pinch-off grows toward the left of the figure until the voltage at the channel is less than 0.5 V. The current is exactly the same as it was in Figure 8.13A (remember we cannot shut off the channel). As we increase the drain voltage still more to 10 V (Figure 8.13C) and 14 V (Figure 8.13D), the pinch off region in the channel keeps on moving to the left, keeping the current constant. If we push the drain voltage further, above 16 V, the whole thing collapses and we have a breakdown condition. I show these four cases, A to D, in the JFET characteristic curves, with dots at V_D = 0.5, 5, 10 and 14 V.

As long as we work in the saturation region, the current is controlled by the voltage at the gate. If I add a sinusoidal signal at the gate, the current follows the same shape as the sinusoidal voltage. I describe the sinusoidal operation in the next chapter.

8.4 The Metal Oxide Semiconductor FET

Another type of transistor, and the most commonly used transistors in integrated circuits (ICs), are MOSFETs. I show the structure and the basic bias of a MOSFET in Figure 8.14.

First let's take a look at the physical construction. In spite of all the boxes and colors, the structure is very similar to the JFET, except that I replaced the upper p-type semiconductor by a metal (in black) separated from the p-type semiconductor (in light blue) by an oxide (in red) thus the name metal-oxide semiconductor field-effect transistor or MOSFET for short. I have also added two highly doped n-type regions, n+ (in dark yellow), at the two ends to make contact to the outside circuit. The MOSFET is, like the JFET, a unipolar device since the moving charges are only electrons (or only holes in an n-type substrate). I explain here the n-type MOSFET, but the same

explanation applies to a p-type MOSFET with the labels changing from p to n and vice versa and reversing all the biases. Notice now that the gate is electrically isolated from the semiconducting channel. The insulating material is usually silicon dioxide, SiO_2. It is a voltage-driven device, that is, there is no current at all from the metallic gate through the oxide to the channel.

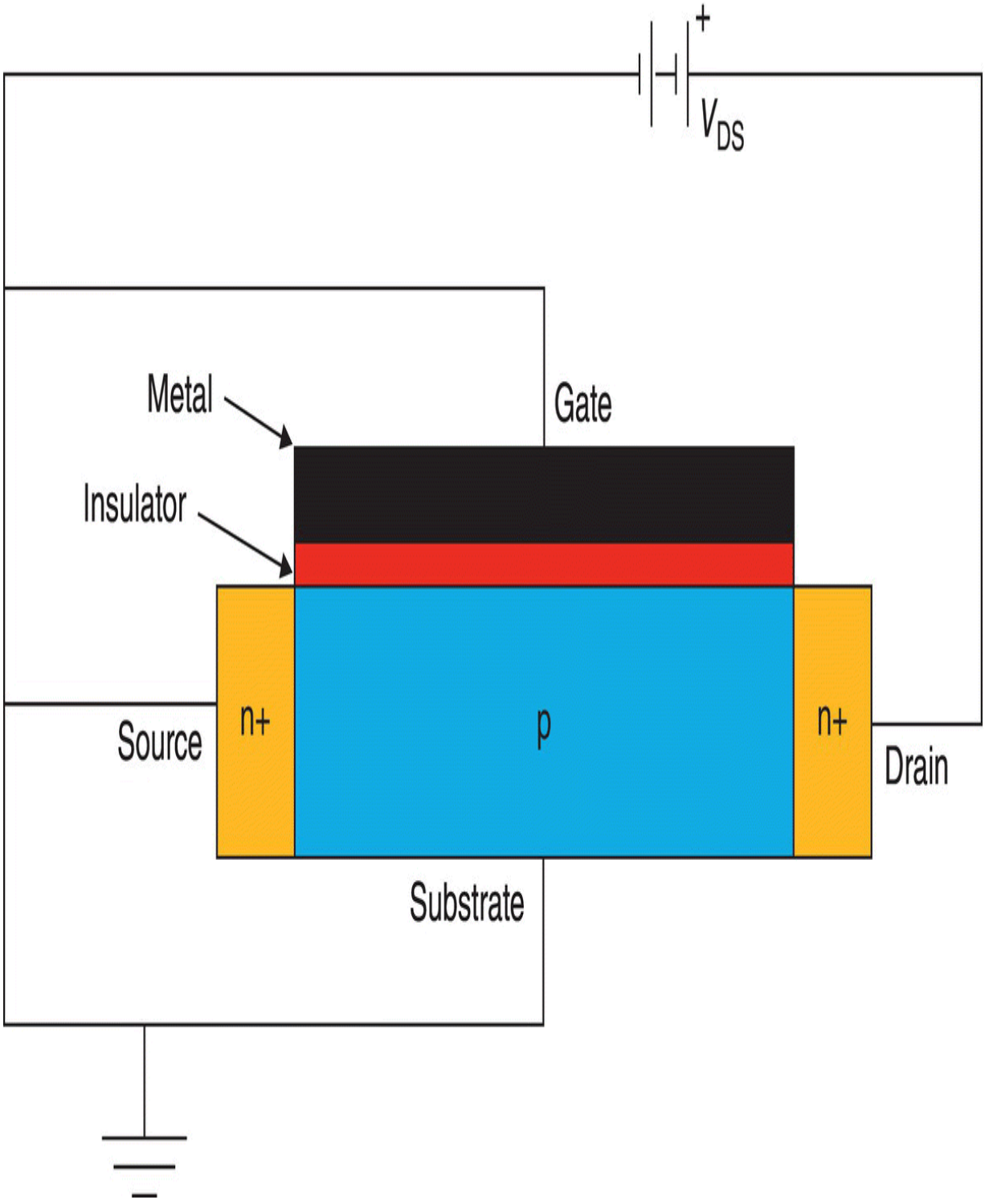


Figure 8.14 In a MOSFET one of the two semiconductor gates in a JFET is replaced by a metallic gate.

In the JFET, the gates are a p-type semiconductor and thus to work

they have to be reversed biased and there is always a residual current in a reversed biased junction, the leakage current. In the MOSFET, the gate is a metal separated from the semiconductor by an insulator. Thus, the voltage at the metal can be either positive or negative and absolutely no current flows through the insulating oxide.

The MOSFET has many advantages over BJTs and JFETs. It is much easier to fabricate, it can be very small, and it has very high input resistance, that is, when the gate voltage is zero or grounded, as I show in Figure 8.14, there is a very high resistance between the source and the drain. Let me also emphasize that what we want to do is to control the current between the source and the drain by changing the voltage at the gate. In Figure 8.14 I have a battery V_{DS} indicating that I am interested in getting current between the source and the drain.

Now consider what happens when I apply a positive voltage at the gate with respect to the source (Figure 8.15).

When I apply a positive voltage between the gate and the source, the positive charges at the metallic gate start attracting negative charges not only from the few available electrons in the p-type material but also from the highly n+-doped semiconductors at the two ends, the drain and the source. As the gate voltage increases, there are more and more electrons attracted to the surface, making the channel more and more conductive. The resistance of the channel decreases and the current between source and drain for a given source to drain voltage, V_{SD}, increases.

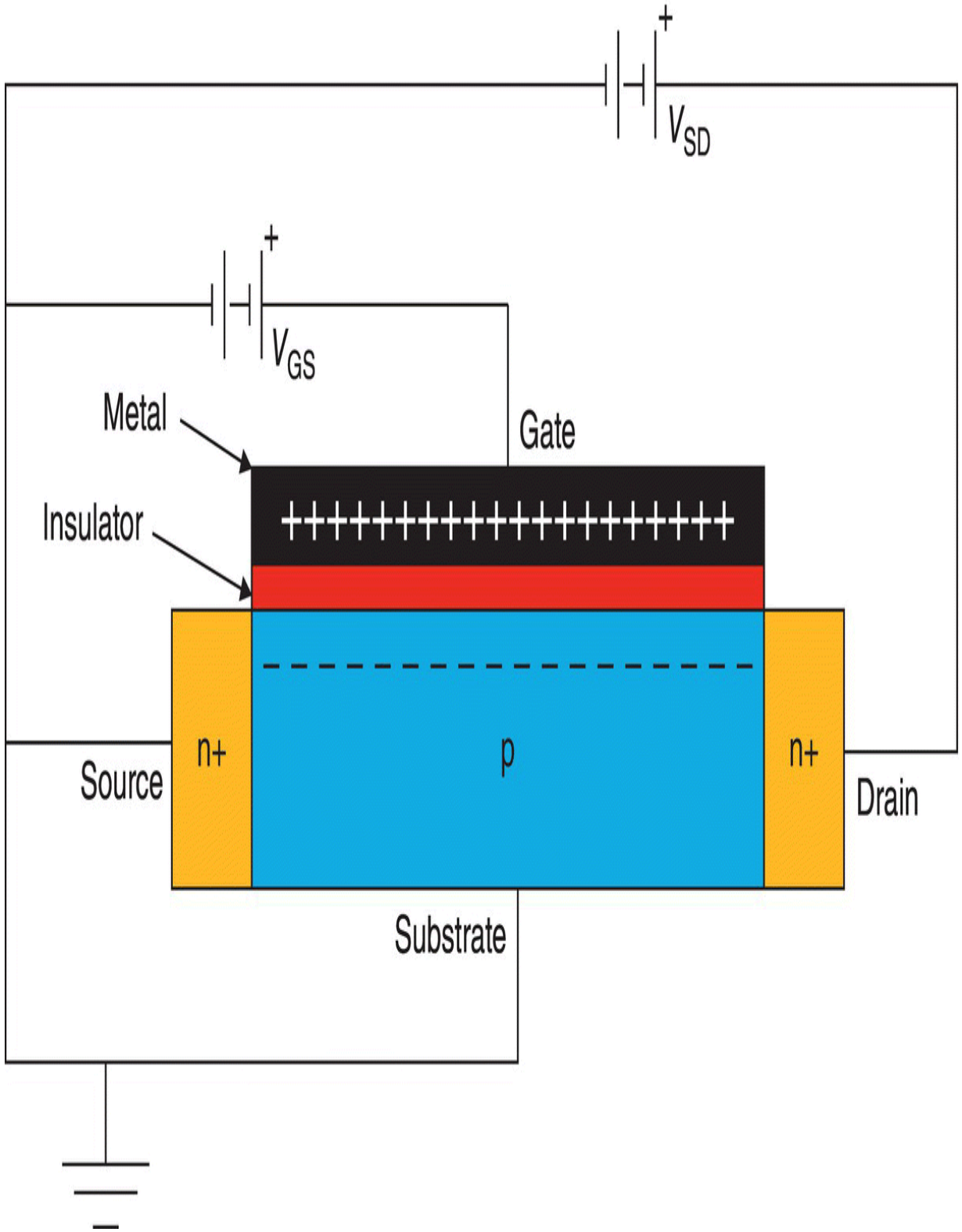

+
V_{SD}
+
V_{GS}
Metal
Gate
Insulator
n+
p
n+
Source
Drain
Substrate

Figure 8.15 If the gate of a p-type MOSFET is positive, electrons are attracted to the interphase between the oxide and the channel, and create a conductive path between the source and the drain.

Let me go into a little more detail, similar to what I did for the JFET in the previous section. Notice that the voltage between the source and the gate at the left is equal to whatever the voltage V_{GS} is. But at the right, the voltage between the gate and the drain is the gate voltage, $V_{GS,}$ minus the drain voltage, V_{SD}. Therefore, as we saw in the JFET, the conductive channel we create is not uniform.

Let me see if I can show what the electron channel looks like as the drain voltage, V_{SD}, increases and the gate voltage remains the same. I'll use some numbers to help understand the process. Let us say that the gate voltage, V_{GS}, is 2 V. As the V_{SD} increases from zero to 2 V, that is, less than the gate voltage, the current increases linearly because the channel shown in Figure 8.15 acts like a resistor, its value changing depending on the voltage V_{GS}. When the drain to sourced voltage, V_{SD}, equals the gate voltage, V_{SG}, i.e. 2 V, the region near the drain is "pinched-off," that is, there are just enough charges under the oxide at the drain end to permit the current to continue. Why? Because there is the same equilibrium condition that we saw in the JFET (Figure 8.11). The long triangle from drain to source (light green in Figure 8.15) shows the area where there are free electrons, practically none at the drain side and quite a few at the source side. If I increase the drain voltage V_{SD} to 3 V, the pinch-off occurs sooner as shown as a triangle in the middle (darker green). Further increases of the drain voltage move the pinch-off voltage to the left (darkest green). The pinch-off effect is very important because it says that the current increases until the drain voltage is equal to the gate voltage but then remains constant until the drain voltage is large enough to create a breakdown. On the left of Figure 8.17 I show an idealized voltage/current plot of this effect.

As the gate to source voltage increases, more and more electrons are attracted to the channel and the resistance of the channel

decreases, generating more current between the source and the drain, the resistive (quasi linear) region. When the gate and the drain voltages are equal, the channel is pinched-off and, as the drain voltage increases still further, the current remains constant (the saturation region). As the drain voltage increases further and further, at some point the junctions between the substrate and the two n-type contacts will break down and the current increases drastically, probably burning the MOSFET. I show the voltage/current characteristics of an actual MOSFET on the right of Figure 8.17.

As far as the fabrication is concerned, drain and source are interchangeable. The device is symmetrical. The MOSFET shown in Figures 8.14 to 8.16 is called an *enhancement mode* Mosfet, enhancement because as the gate voltage increases, the current increases. With the gate voltage zero, there is no current. We say that the MOSFET is normally off.

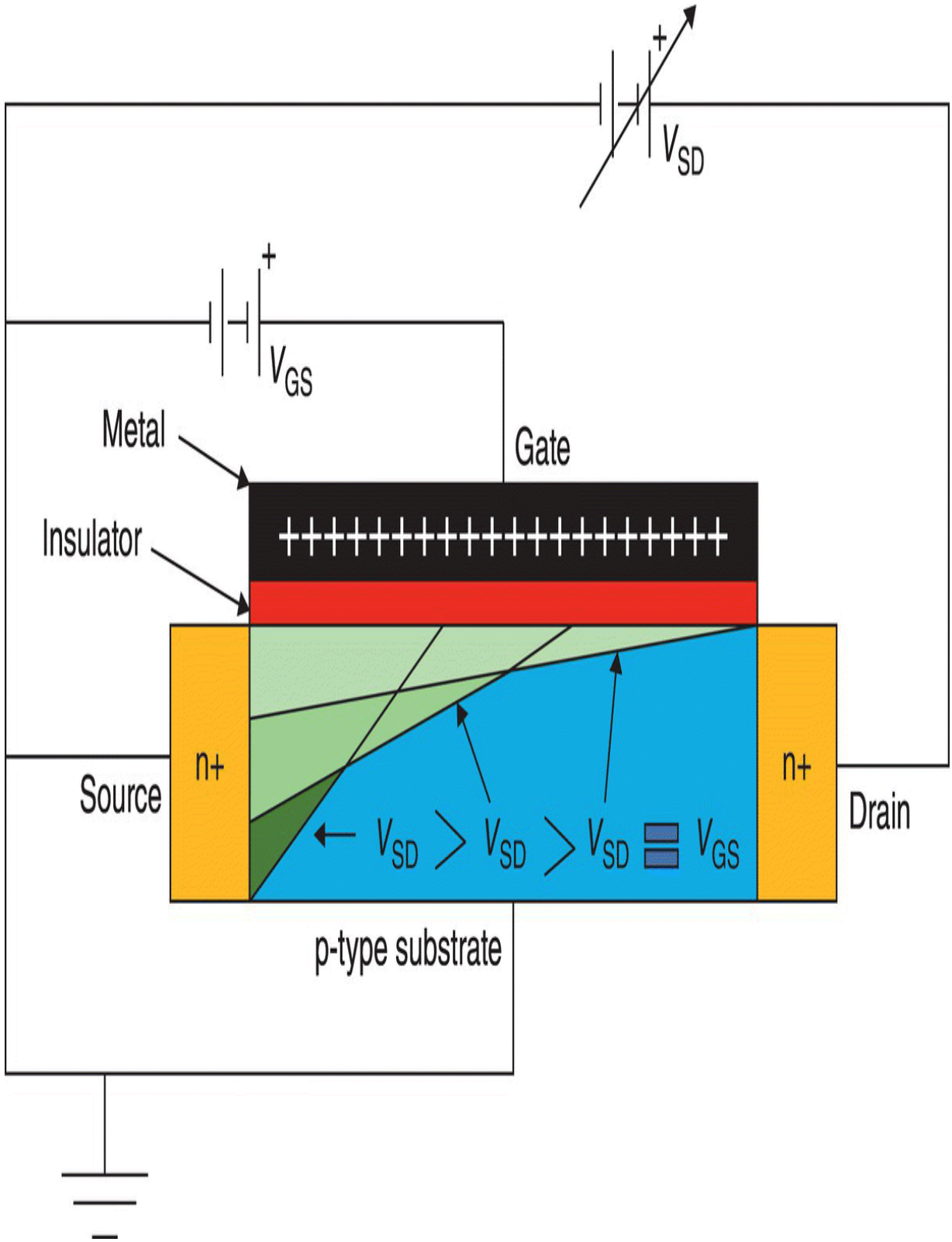

+
V_{SD}
+
V_{GS}
Metal
Gate
Insulator
n+
n+
Source
Drain
$V_{SD} > V_{SD} > V_{SD}$
V_{GS}
p-type substrate

Figure 8.16 A MOSFET showing the region with electrons in the channel under the oxide. As the drain voltage increases, the pinch-off region moves towards the source.

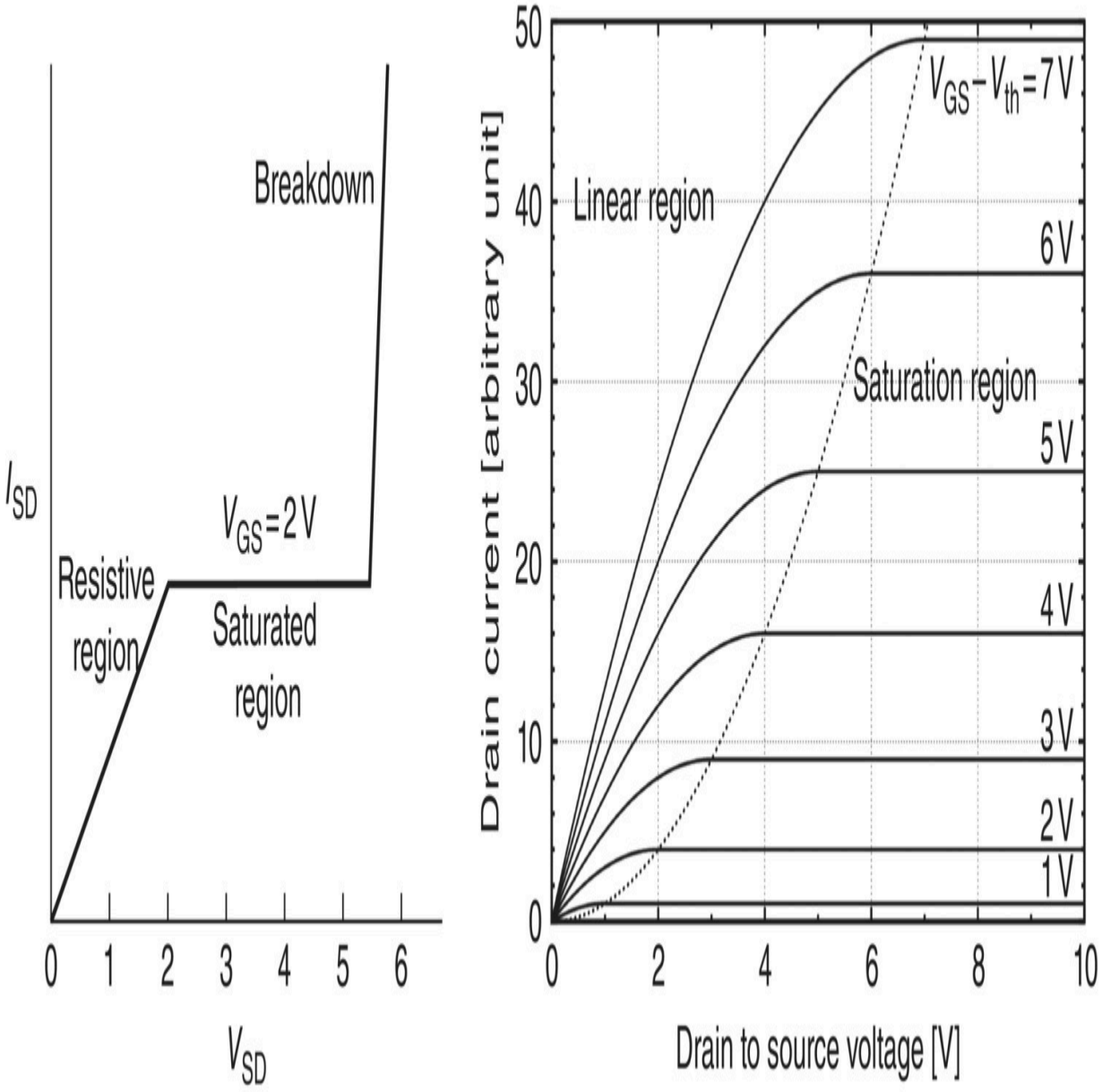


Figure 8.17 Idealized source to drain current as a function of the drain voltage (left) with the gate voltage equal to 2 V and an actual characteristic MOSFET curves (right).

Source: https://en.wikipedia.org/wiki/Current–voltage_characteristic#/media/File:IvsV_mosfet.svg.

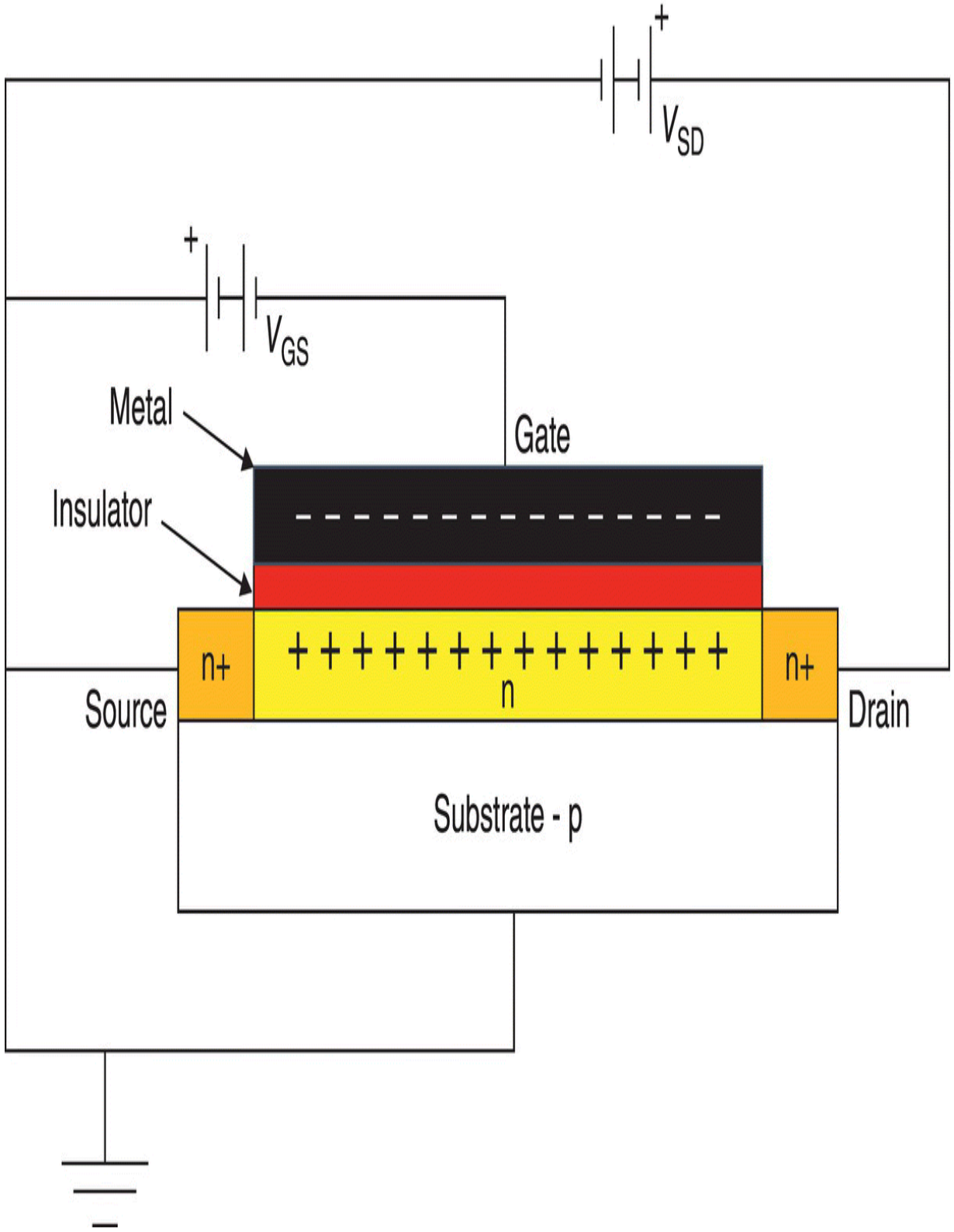

+
V_{SD}
+
V_{GS}
Metal
Gate
Insulator
n+
n+
Source
Drain
n
Substrate - p

Figure 8.18 In a depletion mode MOSFET the channel is made more resistive by attracting charges of the opposite polarity and thus increasing the resistance of the channel.

From the previous sentence you can assume that there is another way of operating the MOSFET. This is the *depletion mode*, and this MOSFET conducts when the gate voltage is zero, i.e. it is ON when the gate voltage is zero and the current decreases as the gate voltage increases. I show the depletion mode MOSFET in Figure 8.18.

Can you see the changes I made? First, I added a thin n-type channel between the substrate from a p-type to an n-type. Second, I reversed the battery V_{GS}. Now, if the gate voltage is zero, there is current from source to drain because the electrons can flow from one n+ contact to the other through the friendly n-type channel, that is, there is continuity of electrons for the current to flow from source to drain. The current is limited by the resistance of the channel. Now suppose, as I show in Figure 8.18, that I apply a negative voltage at the metal gate. The negative charges in the metal plate repel free electrons near the surface and attract holes (positive charges). By reducing the number of electrons in the channel, it becomes more resistive and the current, for a fixed drain voltage, decreases. As we increase the gate voltage the current between source and drain decreases more. Exactly the reverse of the depletion MOSFET. Since the gate is insulated from the channel by an oxide layer, the enhanced MOSFET works with both positive and negative gate voltages.

8.5 Summary and Conclusions

Because the transistor is so fundamental in all modern electronics, I will recapitulate and emphasize the main characteristics. In Figure 8.19 I show a graph of all the transistors I have discussed.

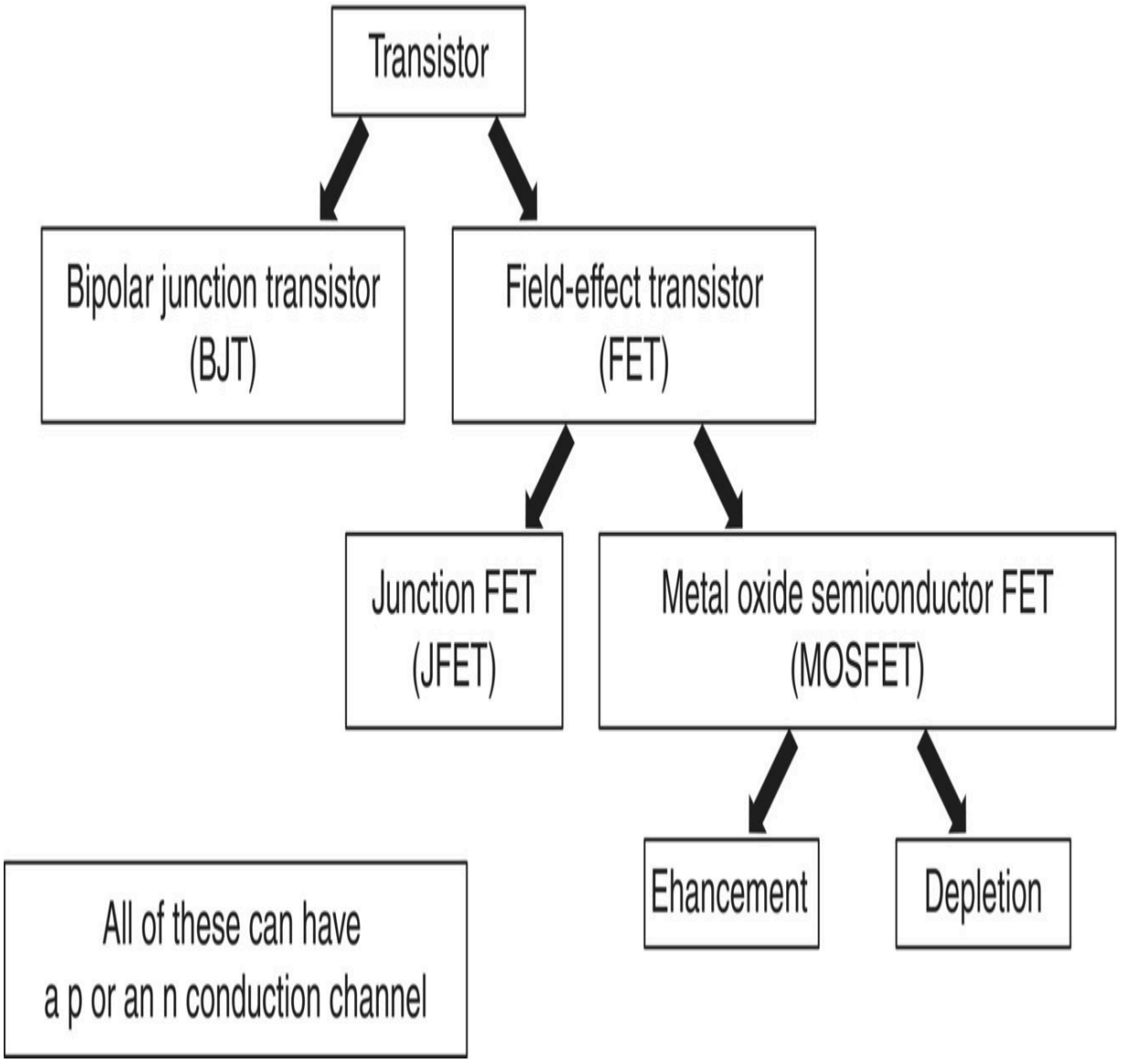


Figure 8.19 The relationships of the variety of transistors discussed in this chapter.

The transistor is any device where one gate controls the current in a different part of the device. There are two basic devices. First, the classic and older bipolar junction transistor (BJT), where one type of semiconductor, the base, is sandwiched between two semiconductor regions of a different type (npn or pnp). The BJT is a current controlled device, that is, the base current controls the current between emitter and collector (see Figure 8.5). It is also a bipolar device because the current is composed of both electrons and holes

that cross the two pn-junctions. The other type of transistor is the FET. In these transistors the current in the device between the source and the drain is controlled by the voltage (not the current) at the gate(s).

FETs can be divided into junction FETs (JFETs) and metal-oxide semiconductor FETs (MOSFETs). MOSFETs can operate in either enhancement mode or depletion mode. All work in basically the same way, with the gate voltage changing the cross-sectional area where the free charges move, therefore increasing or decreasing the resistance and thus the current through the device so it acts like a voltage-controlled resistor. The constant current is obtained by creating a pinch-off region that keeps the current constant as the source to drain voltage increases. All of the FETs are unipolar, that is, the main current, source to drain, is composed only of electrons or only of holes.

The main advantage of MOSFETs is that the gate is insulated from the rest of the circuit, thus presenting an almost infinite input resistance. Another advantage over JFETs is that in JFETs we have to be sure that the two junction gates are reversed biased. In MOSFETs we do not have to worry about that, and the gate voltage can be either positive or negative since the metal is separated from the semiconductor by an insulating dioxide, SiO_2. Therefore, it can work in both the enhancement and depletion modes.

The advantages of MOSFETs are lower power dissipation, smaller size, thus higher density of devices in a given area, and the possibility of building analog and digital circuits side by side. It is also possible to fabricate on and off devices in the same substrate. I will discuss this in coming chapters.

We will also talk about CMOS, Complementary MOS, when we discuss microprocessors and memories (see Chapter 11). These are not "new" devices, but they use both a p-type and an n-type device as a unit, complementing each other. They are very useful because they use less power.

We have discussed the main semiconductor devices, the diode and the transistor. In the next chapters I discuss how we fabricate them (Chapter 10), and how we use them to perform mathematical operations (Chapter 11) and create more advance electronic components (Chapter 12) used in optoelectronics (Chapter 13) and computers (Chapter 14). But now, in the next chapter (Chapter 9), I start discussing how we bias and use these transistors in actual useful circuits.

Appendix 8.1 Punch Trough

A very simple sketch, Figure 8.20, explains how we get punch-through that results in very high current levels, the breakout regions.

On the left we have the band diagram of a npn-transistor with the shared Fermi level. As we increase the reversed bias of the collector to base diode (on the right), the forward voltage encroaches more into the transition region and at some point, the lowest curve on the right, the transition region all but disappears, shorting two n-type regions with a large voltage. The current increases drastically, no longer controlled by the base. A similar process occurs with FETs.

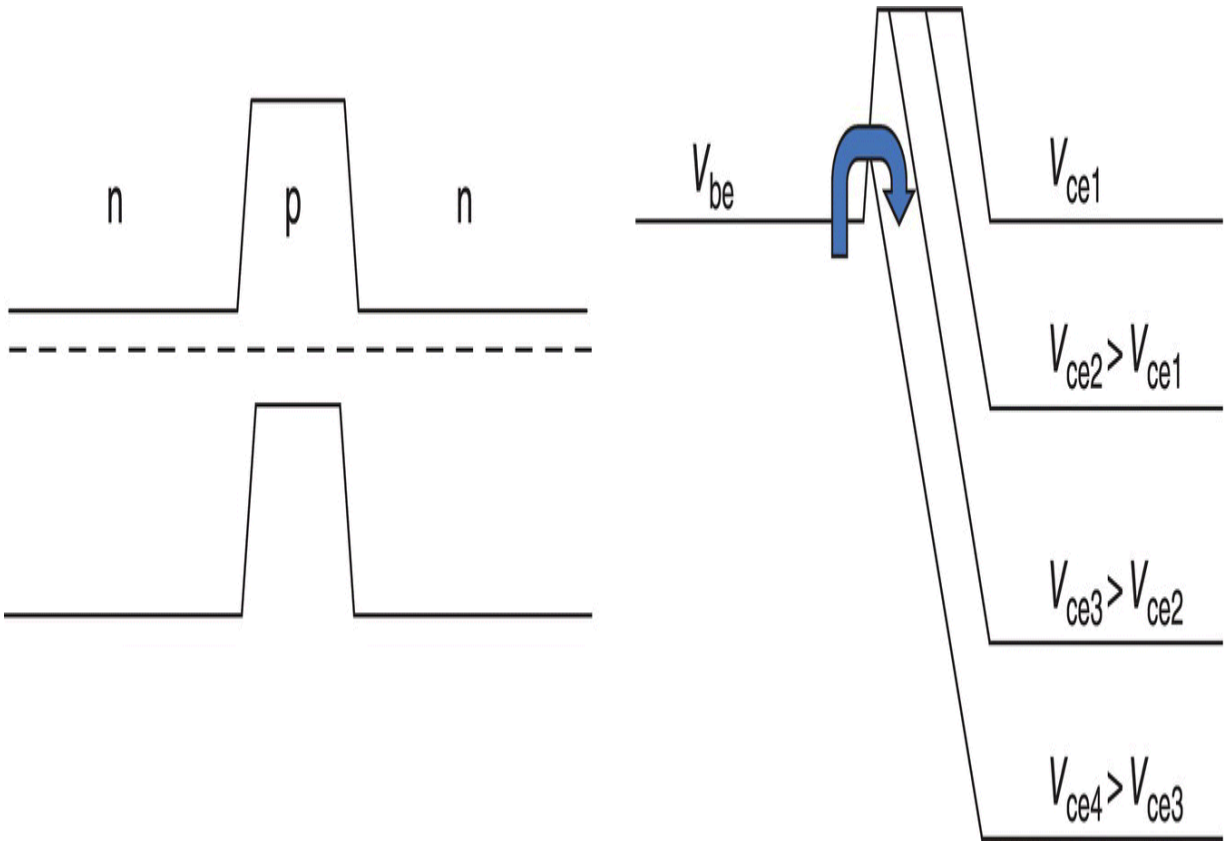


Figure 8.20 The energy bands in an npn-transistor (left) and what happens to the bands as the base to collector voltage increases.

9
Transistor Biasing Circuits

OBJECTIVES OF THIS CHAPTER

I have already explained (almost) all the components that we need to build very complicated electronic devices using semiconductors. As I said before, Forbes has calculated that as of May 2014 there were 2 913 276 327 576 980 000 000 transistors shipped. This is about 3×10^{21}. Just for comparison the human body has between 30 and 70 trillion cells or at least 3×10^{13}. There are100 million more transistors in the world that there are cells in the human body.

In this chapter I explain how we use these transistors, how we bias them, so they generate working and useful circuits, and how we stabilize them as the temperature, currents, and applications vary. We use the term "bias" or "biasing" to indicate the process of connecting the transistor to electrical sources and other components so the transistor operates with optimal performance. I concentrate in this chapter on the basic operation of a single transistor, but I will use these concepts in the following chapters to understand how semiconductor-based devices work. I will also discuss the operational amplifier, a "single component", which is more stable and easier to insert in a circuit without much need for design and calculations.

9.1 Introduction

One thing I want to remind you of before I discuss the different transistor biasing methods is that the transistor parameters are not

constant. The characteristics of transistors change drastically with temperature and current (remember what I said in Chapter 3). We saw this when I talked about the leakage current, I_{CEO}, in Figure 8.8. Although the collector current, I_C, is supposed to be constant as a function of the base current, I_B, the leakage current increases as a function of the collector to emitter voltage, V_{CE}. As we turn on the devices, they warm up, as you may notice when you have a laptop on your lap for a while. The change in temperature changes the gain of the transistor and also, in much less of an effect, the value of all the other components, such as resistors. We use different biasing methods to stabilize the performance of the circuits to overcome these changes. Even transistors from the same supplier and the same ID number have substantial different beta values.

There are several ways you can bias a transistor:

Fixed base bias. This involves a very simple circuit, but it has poor stability. We use this fixed base method in switching circuits because they are simple, have fewer components, and the transistors are only used to jump from on to off, so linearity and stability are not much of a concern.

Collector feedback bias. This method also uses few components and it has better stability than fixed base bias.

Emitter feedback bias. This has much better stability than the other two methods and is more commonly used in analog circuits such as sound amplifiers, operational amplifiers (OpAmps), and delicate and highly accurate measuring instruments. It has the greatest stability.

Let us start with method 3, the most complex one, and then go back to the other two.

9.2 Emitter Feedback Bias

Figure 9.1 shows an emitter feedback bias circuit under DC conditions only. In the next section I cover the AC, sinusoidal

current, performance of the same emitter feedback bias circuit.

Before I start calculating some values and explaining in greater detail how it works, I will qualitatively explain why this is the preferred method to bias a transistor when we need good or high stabilization. Let's see if the flow diagram in Figure 9.2 helps. I will ask you to go back and forth between Figures 9.1 and 9.2 using the numbers on the right of Figure 9.2 to identify the explanations.

Let's start following the steps in Figure 9.2.

Consider the case in which as I turn the device ON the temperature goes up.

The resistors R_1 and R_2 do not change much. Additionally, they both change in the same direction, that is, they would simultaneously either increase or decrease their resistance, thus the voltage in the middle, the base voltage V_B (Figure 9.1), changes very little compared with the changes in the collector leakage current, I_{CEO}, and the gain β, which both go up quite a bit. β can easily change from 50 to 250, a factor of 5, as the temperature goes from 25 °C to 100 °C. The rule of thumb is that the current I_{CEO} doubles every 10 °C. As you can see, we are not talking about small changes.

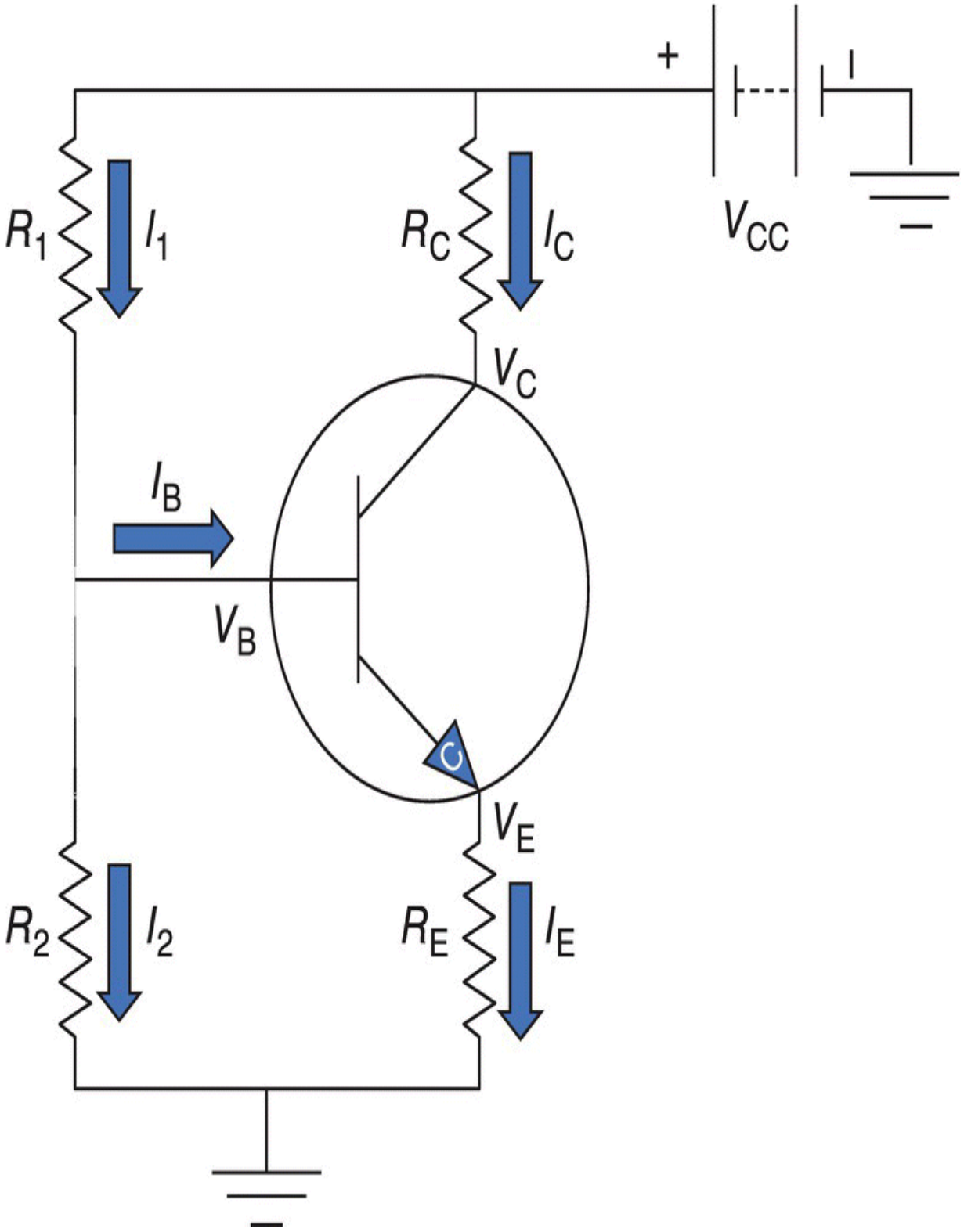


Figure 9.1 The emitter feedback bias circuit has the highest stability as the temperature and currents change.

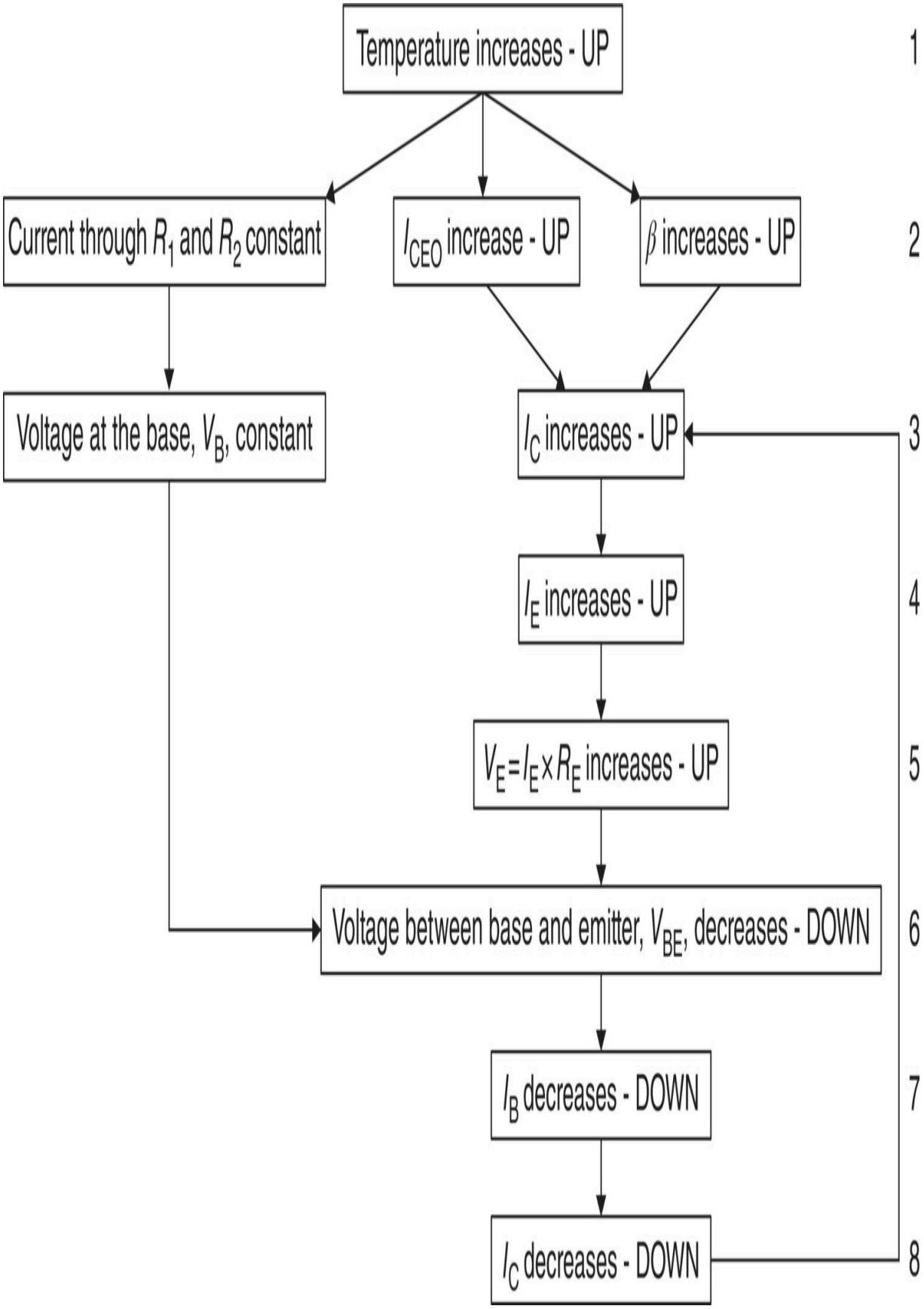

Temperature increases - UP
1
Current through R_1 and R_2 constant
I_{CEO} increase - UP
β increases - UP
2
Voltage at the base, V_B, constant
I_C increases - UP
3
I_E increases - UP
4
$V_E = I_E \times R_E$ increases - UP
5
Voltage between base and emitter, V_{BE}, decreases - DOWN
6
I_B decreases - DOWN
7
I_C decreases - DOWN
8

Figure 9.2 This flow diagram shows how the emitter negative feedback stabilizes the transistor operation. As I_C tries to increase (line 3), I_B and therefore I_C (lines 7 and 8) try to decrease, thus both currents stay constant.

As I_{CEO} and β go up the collector current, I_C, "tries" to also go up, while the base voltage, V_B, defined by the ratio of R_1 and R_2, remains almost unchanged.

Since the emitter current, I_E, is approximately equal to the collector current, it also "tends" to go up by the same amount as the collector current.

The voltage V_E across the emitter resistor R_E also "tends" to go up since it is the product of the current through the emitter, I_E, times the emitter resistance, R_E, which does not change or changes very little.

Since the base voltage, V_B, does not change and the voltage V_E "tends" to go up, the voltage between the base and the emitter, V_{BE}, will "tend" to go down.

If the voltage V_{BE} "tends" to go down, so does the base current, I_B.

Therefore, the collector current, $I_C = \beta I_B$, "tends" to decrease. This is exactly the opposite of the situation in step 3.

The increase in step 3 and the opposing decrease in step 8 are, by definition, negative feedback that keeps the DC conditions constant. If you read point 8 again you should now understand why I use the word "tend" in quotation marks. The current I_C does not increase or decrease. It is being push in both directions.

So, now that we understand how this emitter bias circuit is supposed to work, let's go back to Figure 9.1 and decide what values the resistors need to have to operate this transistor circuit in this mode. This is not that complicated. Let us go one step at a time. Review Figure 9.1.

There is one DC voltage source, V_{CC}, which provides the constant currents. One path goes through the resistors R_1 and R_2 on the left, and the other through the resistor R_C, the transistor, and the resistor R_E on the right. The relationships between these currents are

$$I_E = I_C + I_B \quad \text{and} \quad I_1 = I_2 + I_B \tag{9.1}$$

Since I_B is very small compared to all the other currents, I can approximate the relationships in (9.1) above by writing

$$I_E \approx I_C \quad \text{and} \quad I_1 \approx I_2 \tag{9.2}$$

where the wiggly equals sign, ≈, means approximately equal.

Since the purpose of this book is not to teach you how to design circuits, but to give you an understanding of how circuits work, in Figure 9.3 I tell you the resistance values I select and I will explain why I chose these values.

Figure 9.3 is the same as Figure 9.1, with values added for the four resistors and the bias voltage, V_{CC}. Also look at Figure 9.4. It is the same as the right-hand side of Figure 8.8 but I have added a line, called the load line, and a point, Q, in the middle of the line. I will go back and forth between Figures 9.3 and 9.4.

This transistor has a gain of about 50, $\beta = 50$, at room temperature. Look at Figure 9.4. When voltage V_{CE} is 2 V (the short vertical dotted line), the collector current changes from about 2 to 3mA and the base current changes from 40 to 60 μA. Therefore we can say that β is

$$\beta = \frac{\Delta I_C}{\Delta I_B} = \frac{(3-2)\text{mA}}{(60-40)\mu\text{A}} = \frac{1000\mu\text{A}}{20\mu\text{A}} = 50 \tag{9.3}$$

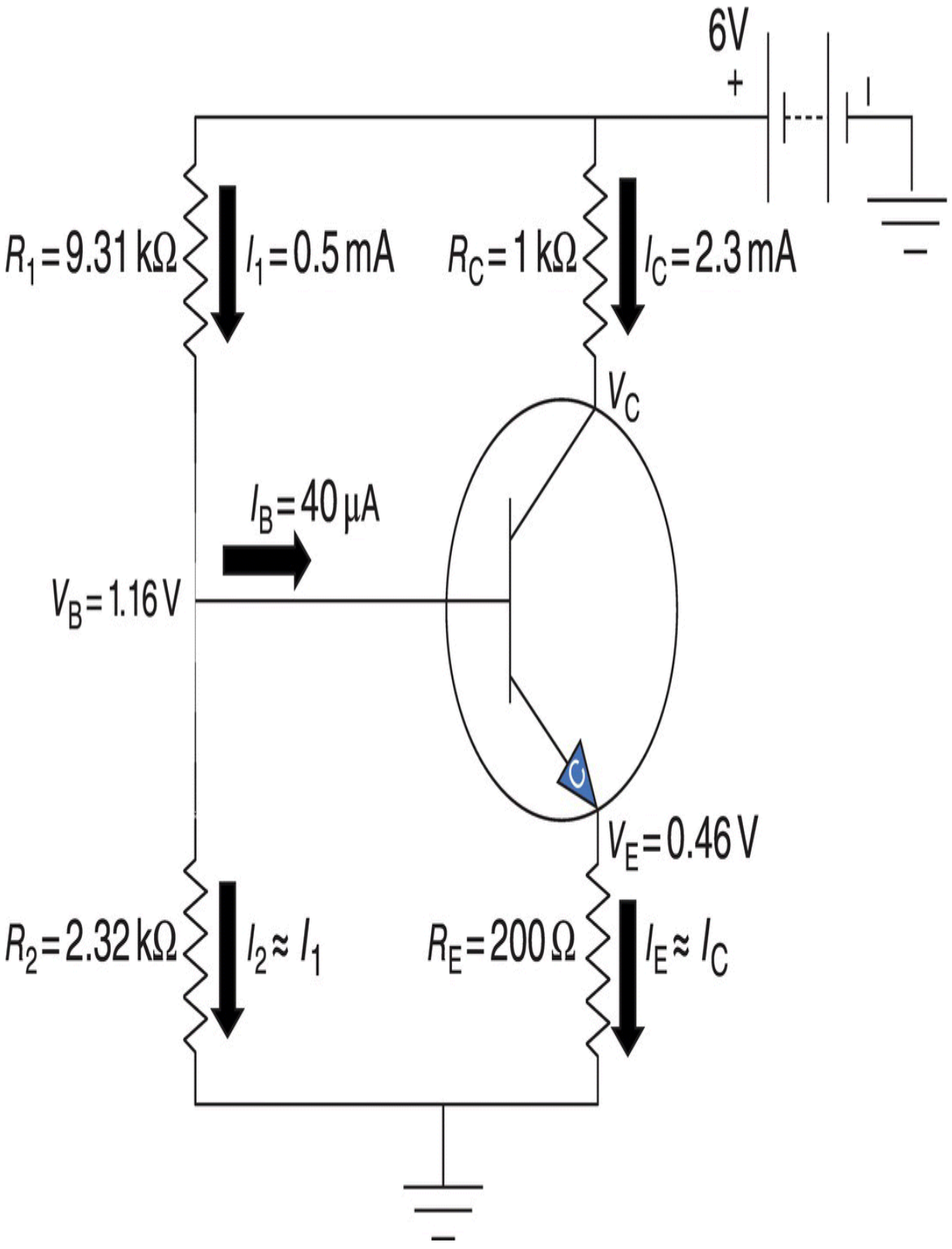

6V
+
$R_1 = 9.31\,k\Omega$
$I_1 = 0.5\,mA$
$R_C = 1\,k\Omega$
$I_C = 2.3\,mA$
V_C
$I_B = 40\,\mu A$
$V_B = 1.16\,V$
C
$V_E = 0.46\,V$
$R_2 = 2.32\,k\Omega$
$I_2 \approx I_1$
$R_E = 200\,\Omega$
$I_E \approx I_C$

Figure 9.3 Emitter feedback bias circuit with the resistor values we need to operate and stabilize the transistor.

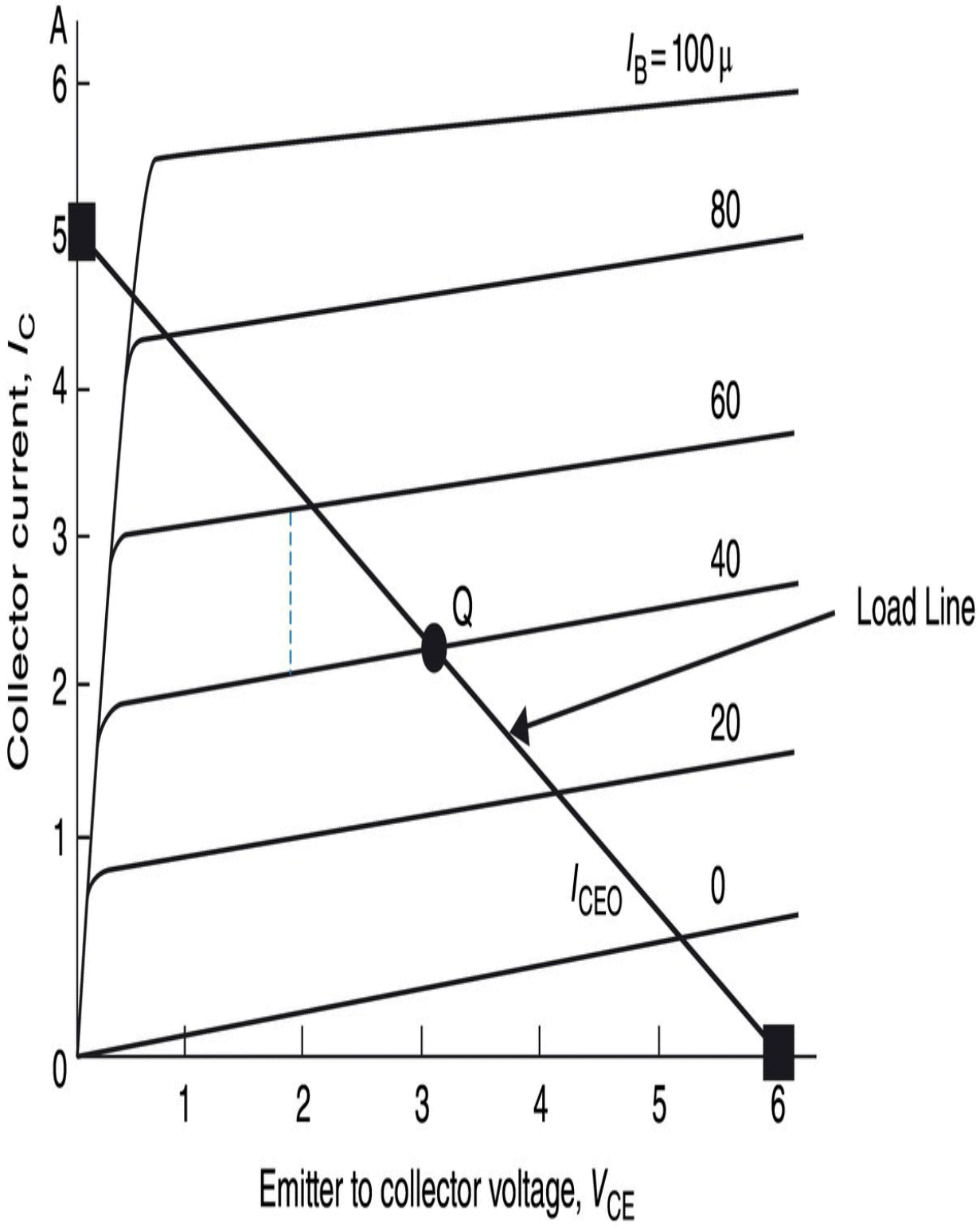


Figure 9.4 The load line that determines the output voltage–current relation of a properly biased transistor and the desired DC operational point, the Q-point, is superimposed on the transistor characteristics curves of Figure 8.8.

Look at the transistor characteristic curves in Figure 9.4. First, we know that when the collector current I_C is zero, there is no voltage drop across the resistors R_C or R_E, thus the voltage across the transistor, V_{CE}, must be equal to the battery voltage, i.e. 6 V. This is the square point on the *x* axis, where $I_C = 0$ and $V_{CE} = 6$ V. When the voltage across the transistor is zero, the battery voltage, V_{CC}, is divided between resistors R_C and R_E, and that is the maximum current I can get given the specific values of the resistors. I want the maximum collector current, I_C, not to exceed 5 mA because I don't want the current to increase above the transistor's linear region. This is the square point on the *y* axis for the vertical line of the characteristic curves, Therefore the collector current, I_C, which is equal to $V_{CC}/(R_C + R_E)$, must be equal to 5 mA. These two points define a line that we call the *load line*, which determines graphically what the voltage across the transistor is as the collector current, I_C, increases from 0 to its maximum value of 5 mA. In the middle of the load line there is a dot at the intersection of the load line and the 40 μA line. This point is called the *quiescent point*, or the Q-point. This is the desired operating conditions of the transistor when there is only a constant, DC, voltage. We want to operate the transistor at this Q-point at the middle of the load line so that when we add a sinusoidal input current (see the next section) the output current can swing between 0.4 and 4.2 mA without falling off the linear portion of the characteristic curves. The Q-point shows the operation of the transistor when there is no signal added, that is, only DC, all quiet, no voltage or current changes.

Now that we know we want 5 mA when the voltage across the transistor is zero, the voltage V_{CC} must be dropped across the resistors R_C and R_E, therefore

$$R_E + R_C = \frac{V_{CC}}{I_C} = \frac{6\,\mathrm{V}}{0.005\,\mathrm{A}} = 1200\,\Omega \tag{9.4}$$

I select R_E = 200 Ω and R_C = 1000 Ω (you'll see why I want R_C >> R_E later).

Let's now look at the left side of Figure 9.3. We want the current through R_1 and R_2 to be at least 10 times larger than I_B, and we have chosen I_B = 40 µA (the Q-point on the load line). Let's then select the current I_1 = 500 µA, for example. Therefore, the sum of the two resistances is

$$R_1 + R_2 = \frac{V_{CC}}{I_1} = \frac{6\,\mathrm{V}}{0.0005\,\mathrm{A}} = 12\ 000\,\Omega = 12\,k\Omega \tag{9.5}$$

Now we have to divide this total resistance between R_1 and R_2. By looking at where I selected the quiescent point, we see that the current I_C = 2.3 mA, therefore the emitter voltage, V_E, across the resistor R_E must be

$$V_E = I_E R_E \approx I_C R_E = 2.3\,mA \times 200\ ' \quad = 0.46\,\mathrm{V} \tag{9.6}$$

The voltage at the base, V_B, is equal to the voltage across R_E plus the turn-on voltage of the transistor (remember the detector turn-on voltage, Figure 5.9 in Section 5.2), which is typically around 0.7 V. Thus, voltage V_B should be:

$$V_B = V_E + V_T = 0.46 + 0.7 = 1.16\,\mathrm{V} \tag{9.7}$$

Now that we know the voltage and the current I want across R_2, we can calculate the resistance:

$$R_2 = \frac{1.16\,\mathrm{V}}{0.5\,\mathrm{mA}} = 2.32\,\mathrm{k\Omega} \tag{9.8}$$

Therefore from Eqs. (9.5) and (9.8)

$$R_1 = 12\ 000 - 2320 = 9.36\,\mathrm{k\Omega} \tag{9.9}$$

Using the closest standard resistor values, I select the values of 2.32 kΩ for R_2 and a 9.31 kΩ for R_1. Take another look at Figure 9.3. We have designed a stable transistor circuit. When we turn this transistor on and it starts warming up, the currents in and out of the transistor change very little.

9.3 Sinusoidal Operation of a Transistor with Emitter Bias

Now that we have biased the transistor circuit so that it is stable, let's see how we can introduce a sinusoidal signal. Take a look at Figure 9.5.

In Figure 9.5, I have added the following components to Figure 9.1:

A sinusoidal input signal source, V_{in}, on the left, connected to the transistor base with a capacitor C_1.

An output resistor, R_O, representing a load to the amplifier, connected to the collector terminal with another capacitor, C_2. The output resistance R_O represents whatever other device is connected to the output, which could be a speaker or another amplifying stage.

Another capacitor, C_3, across the emitter resistor R_E.

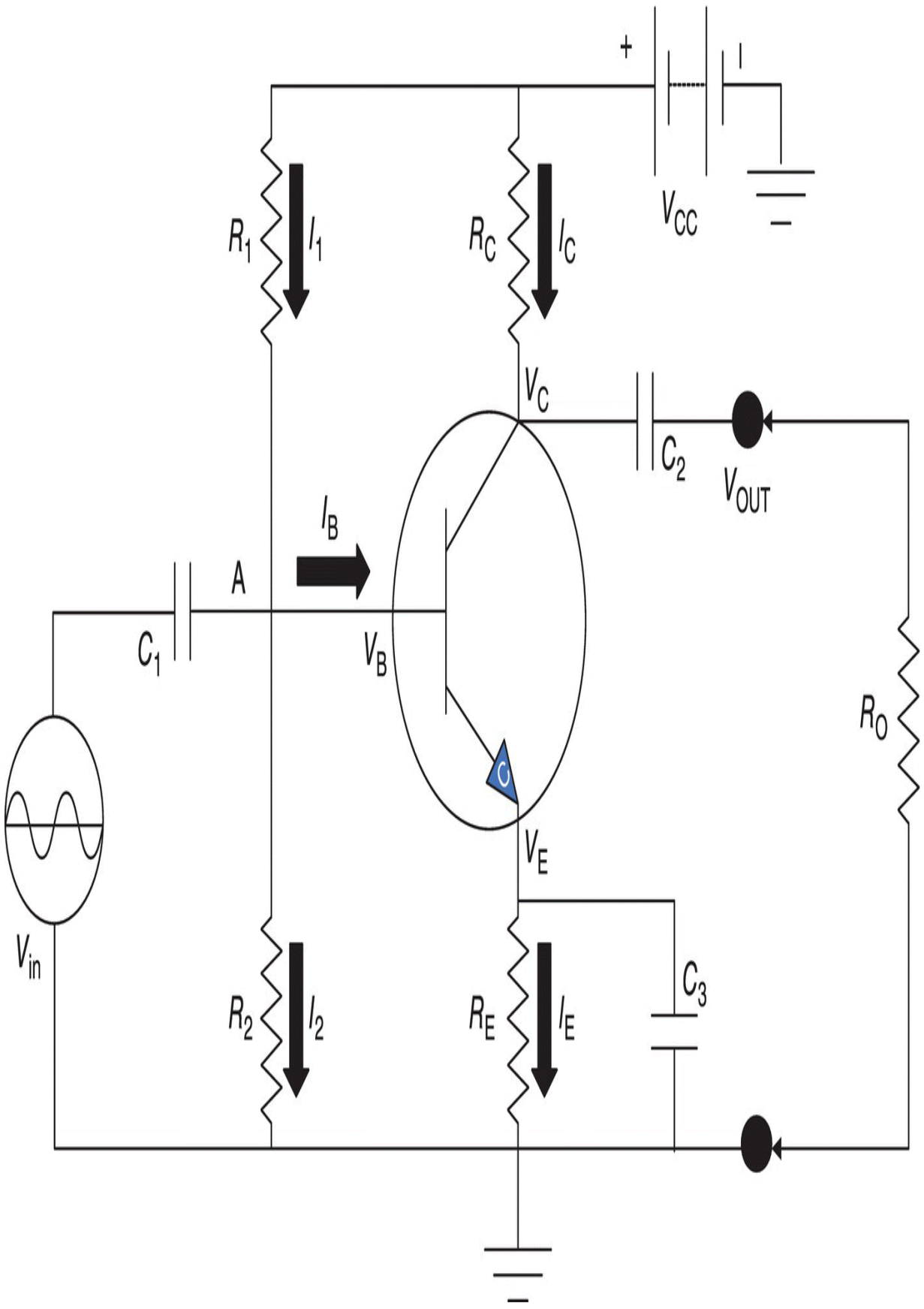

+
V_{CC}
R_1
I_1
R_C
I_C
V_C
C_2
V_{OUT}
I_B
A
C_1
V_B
R_O
V_E
V_{in}
R_2
I_2
R_E
I_E
C_3

Figure 9.5 By adding a sinusoidal signal using capacitors we can modulate the output voltage without changing the circuit's DC conditions.

One thing you should realize right away is that the DC conditions we developed in the previous section have not changed at all. All of these added elements are connected by capacitors which, as far as the DC currents are concerned, are open circuits. The AC currents and voltages are superimposed on the DC currents. The emitter resistor, R_E, as far as the sinusoidal signal is concerned, is shorted, and so is the battery. Therefore, the sinusoidal signal does not see the resistor R_E or the battery at all. I can say then that the circuit above, from a sinusoidal voltage point of view, looks like Figure 9.6.

In Figure 9.6 the capacitors and the battery are shorted. The sinusoidal signal does not see them. I show this by dashing the capacitors and battery, and replacing them by a shorting solid line. Notice also that I have removed the emitter resistor R_E altogether since it is shorted to the sinusoidal voltage by the capacitor C_3.

If you look carefully at Figure 9.6 you will notice that one side of all the three resistors is connected to ground. Also there is a small current, I_B, going into the transistor. The input of the transistor can be represented by an input resistor that we call r_b. From the output point of view, the transistor is a device that sends a current to the collector, the output, equal to the base current times the gain of the transistor. I can then replace Figure 9.6 by Figure 9.7. The transistor is now represented by a box (dotted lines) with an input resistance, r_b, and an output current, βi_b. (I have changed the current symbol from a capital to a lower case letter to indicate that this is the sinusoidal, variable, current, which is a function of time. The standard notation r_b changes with temperature and operation.)

The input sinusoidal voltage sees three resistors is parallel, R_1, R_2 and the input resistance r_b. The transistor input sees a voltage equal to the sinusoidal input voltage, V_{in}. In fact, all three resistors see the

same input voltage. Therefore, the base current, i_b, is equal to the input voltage divided by the transistor input resistance, r_b. So, what is the value of the input resistance? Usually it is very small, in the order of 10 Ω (it can go from 4 or 5 Ω to 1000 Ω: its value is given by the manufacturer). Let us take a look again at the transistor characteristic curves. Figure 9.8 shows the same characteristic curves as in Figure 9.4, but with the AC signals superimposed over the load line.

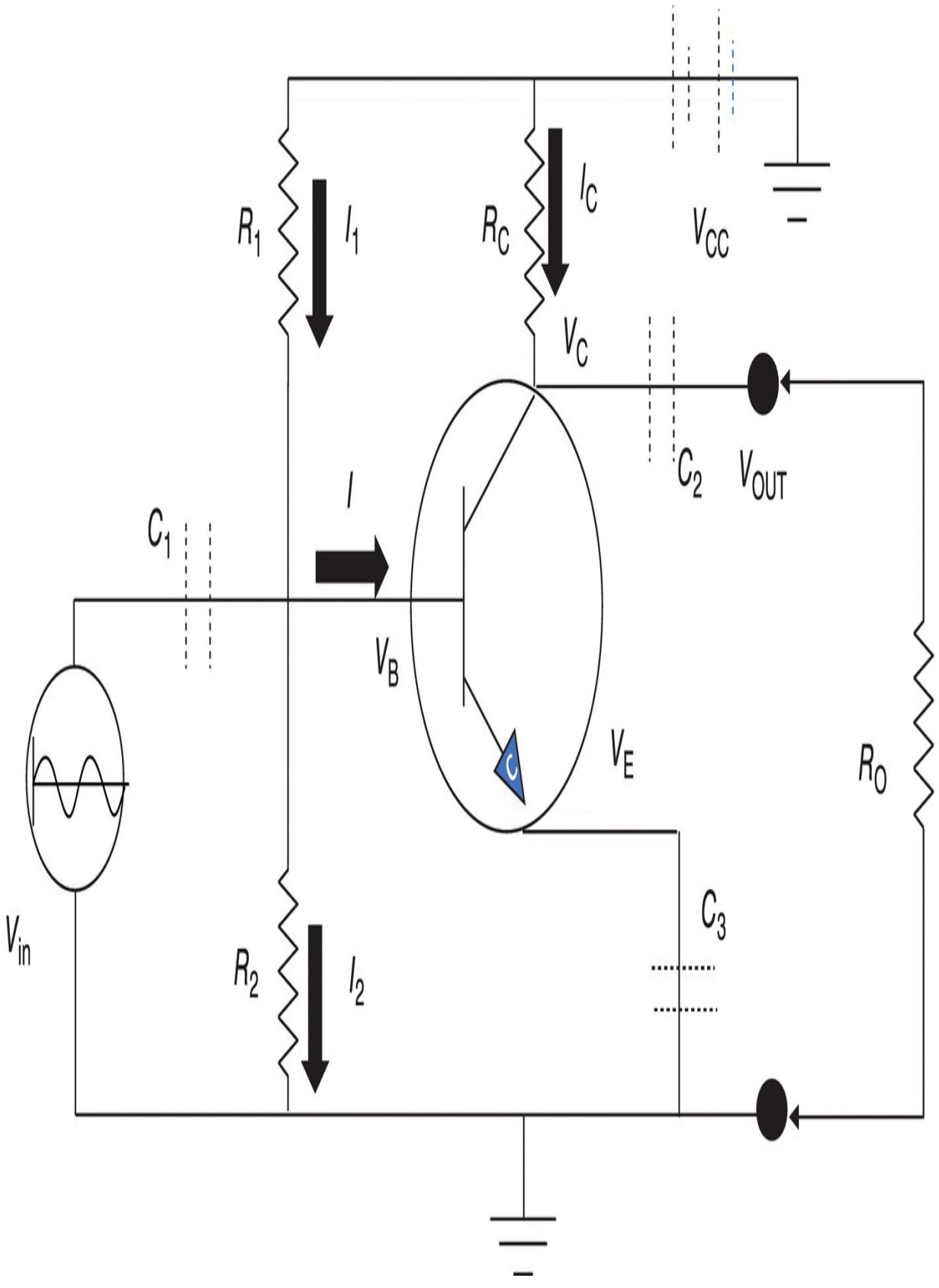

R_1
I_1
R_C
I_C
V_{CC}
V_C
C_2
V_{OUT}
I
C_1
V_B
V_E
R_O
V_{in}
R_2
I_2
C_3

Figure 9.6 From a sinusoidal source point of view the capacitors and the battery are shorted.

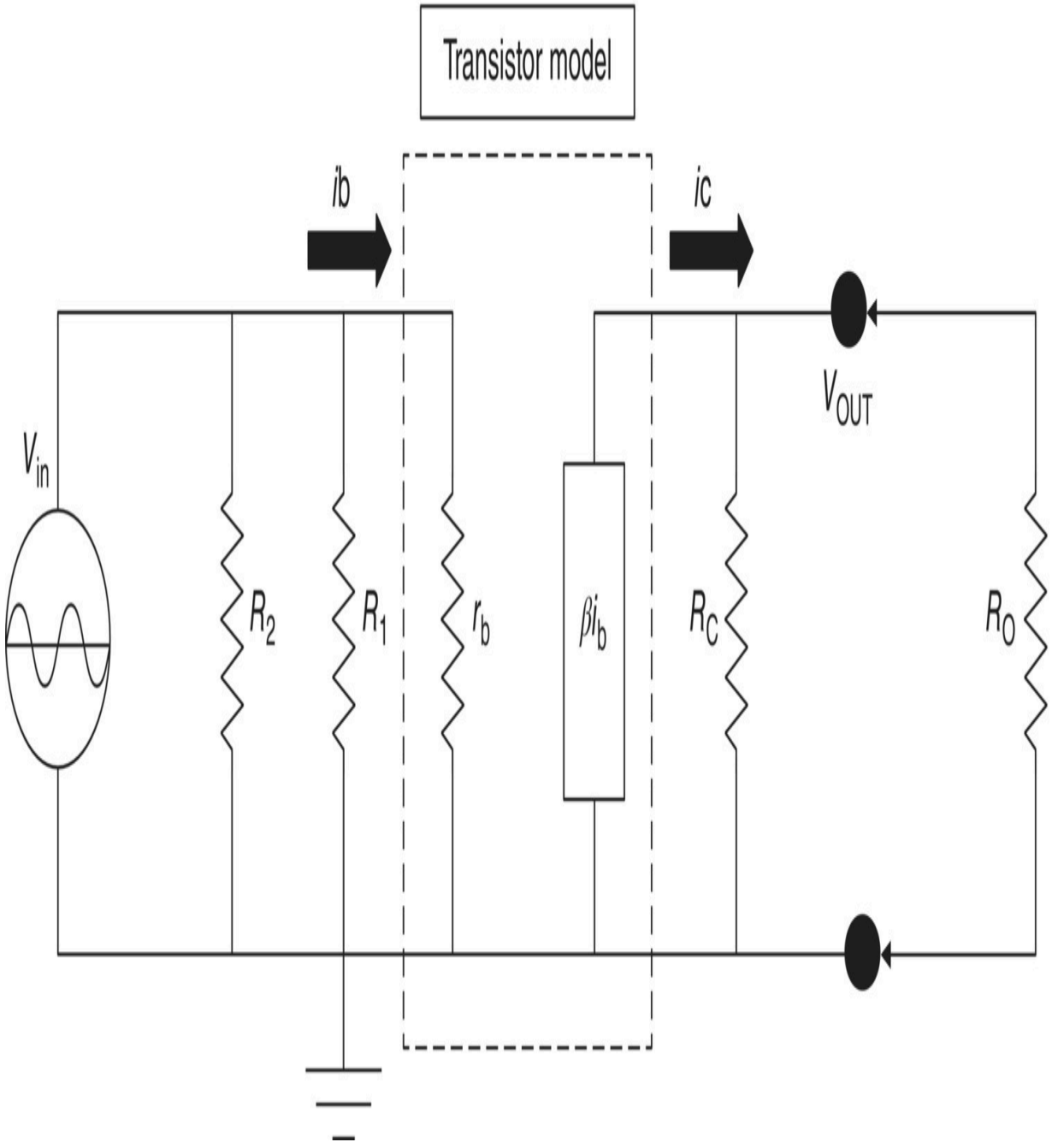


Figure 9.7 The AC equivalent circuit of a transistor consisting of an input resistance and an output current source whose value is the input current, I_B, times the current gain, β.

We want the sinusoidal signal to be restricted to stay in the linear

region, which is basically from 0 to 80 μA, that is, 40 μA from the center Q-point in both directions (the diagonal sinusoidal wave going from 0 to 80 μA). If the input signal is larger than 40 μA, we are going to exceed the linear operating region and distort both ends of the sinusoidal signal. So, since the transistor input resistance is 10 Ω in our case, the input voltage should be no larger than

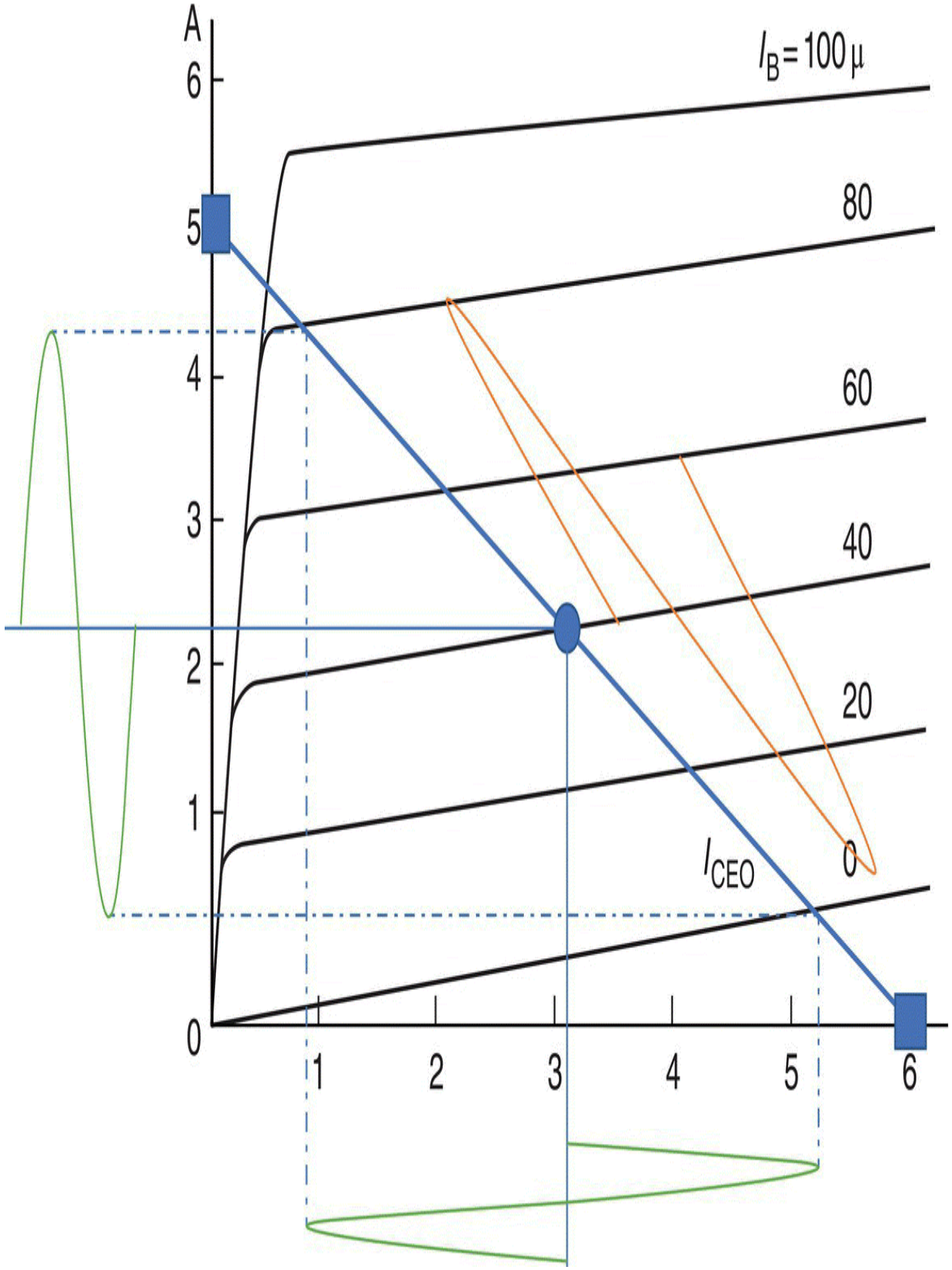

A
6
5
4
3
2
1
0
$I_B = 100\,\mu$
80
60
40
20
0
I_{CEO}
1
2
3
4
5
6

Figure 9.8 We can superimpose the sinusoidal signals on the transistor characteristic curves showing that a sinusoidal change in the base results in a much larger sinusoidal output current and voltage in the output.

$$v_{in} = 40\,\mu A \times 10\,\Omega = 400\,\mu V \tag{9.10}$$

If this is the case, the output current is

$$i_C = \beta i_B = 60 \times 40\,\mu A = 2.4\,mA \tag{9.11}$$

which is what the curves tell you anyway. The collector current, i_C, goes from about 0.5 mA to 4.3 mA (look at the vertical sinusoidal signal at the left of Figure 9.8), or a sinusoidal current of 1.9 mA.

Back to Figure 9.7. The current i_C is divided between the resistor R_C and the output resistor R_O. Let us assume that the output resistance is 25 Ω, small compared to the resistance R_C, which I have chosen to be 1000 Ω Then the majority of the current goes through the output resistance. The output voltage v_{out} is

$$v_{out} = i_C R_O = 200\,\mu A \times 25\,\Omega = 5\,mV \tag{9.12}$$

If the output resistance, R_O, is very large, the voltage is limited by the collector resistance, R_C, which is 1000 Ω and thus the maximum output voltage in this case is

$$v_{out\,max} \approx i_C R_O = 200\,\mu A \times 1000\,\Omega = 0.2\,V \tag{9.13}$$

This is very interesting. Although the current gain of the transistor, β, is constant at a given temperature, the voltage gain depends very much on the characteristic of the circuit that the transistor is trying to drive. If we look at the ratio of v_{out}/v_{in}, we find that the voltage gain can go from

$$\text{voltage gain} = \frac{v_{\text{out}}}{v_{\text{in}}} = \frac{5\,\text{mV}}{0.4\,\text{mV}} = 12.5 \tag{9.14}$$

for a very low output resistance up to

$$\text{voltage gain} = \frac{200\,\text{mV}}{0.4\,\text{mV}} = 500 \tag{9.15}$$

for a very high output resistance.

Quite a swing of output voltage gain!

Now you can see why I selected $R_E \ll R_C$. As far as the sinusoidal signals is concerned the resistance R_E does not exist, it is shorted by the capacitor, and the larger I make R_C, the higher the voltage gain. Remember though that the smaller I make R_E, the worse the stability of the circuit.

Because we have done so much in this section, I summarize here what we have done:

Given a power supply voltage (6 V) and the maximum current we can tolerate to keep the signals in the linear region (5 mA), we draw a load line.

We choose an operating point, which we call the Q-point, on the load line in the middle of the linear operating region.

The Q-point tells us what the DC collector current, I_C, is and therefore also I_E, since they are about the same.

For high stability I choose a large R_E and for large gain I choose a large R_C.

I select the current through R_1 and R_2 to be larger than 10 times I_B.

Using the selected current through the emitter resistor and the turn-on voltage of the transistor, I calculate the base voltage, V_E, and

knowing I_e and the current through R_2 I calculate the value of R_2.

The circuit design is complete and the last thing to do is to calculate the maximum sinusoidal input voltage that we can use without running the transistor into the saturation mode.

We have designed a voltage amplifier.

9.4 The Fixed Bias Circuit

Now let me explain the fixed bias circuit, which is much simpler than the emitter feedback circuit. I show the circuit in Figure 9.9. This is obviously a much simpler circuit that the one in Figure 9.1. Here, the DC base current, I_B, is equal to the current I_1, which is equal to the battery voltage V_{CC} minus the voltage between the base and the emitter, V_{BE}, about 0.7 V. So, the base current is equal to

$$I_B = I_1 = \frac{V_{CC} - V_{BE}}{R_1} \tag{9.16}$$

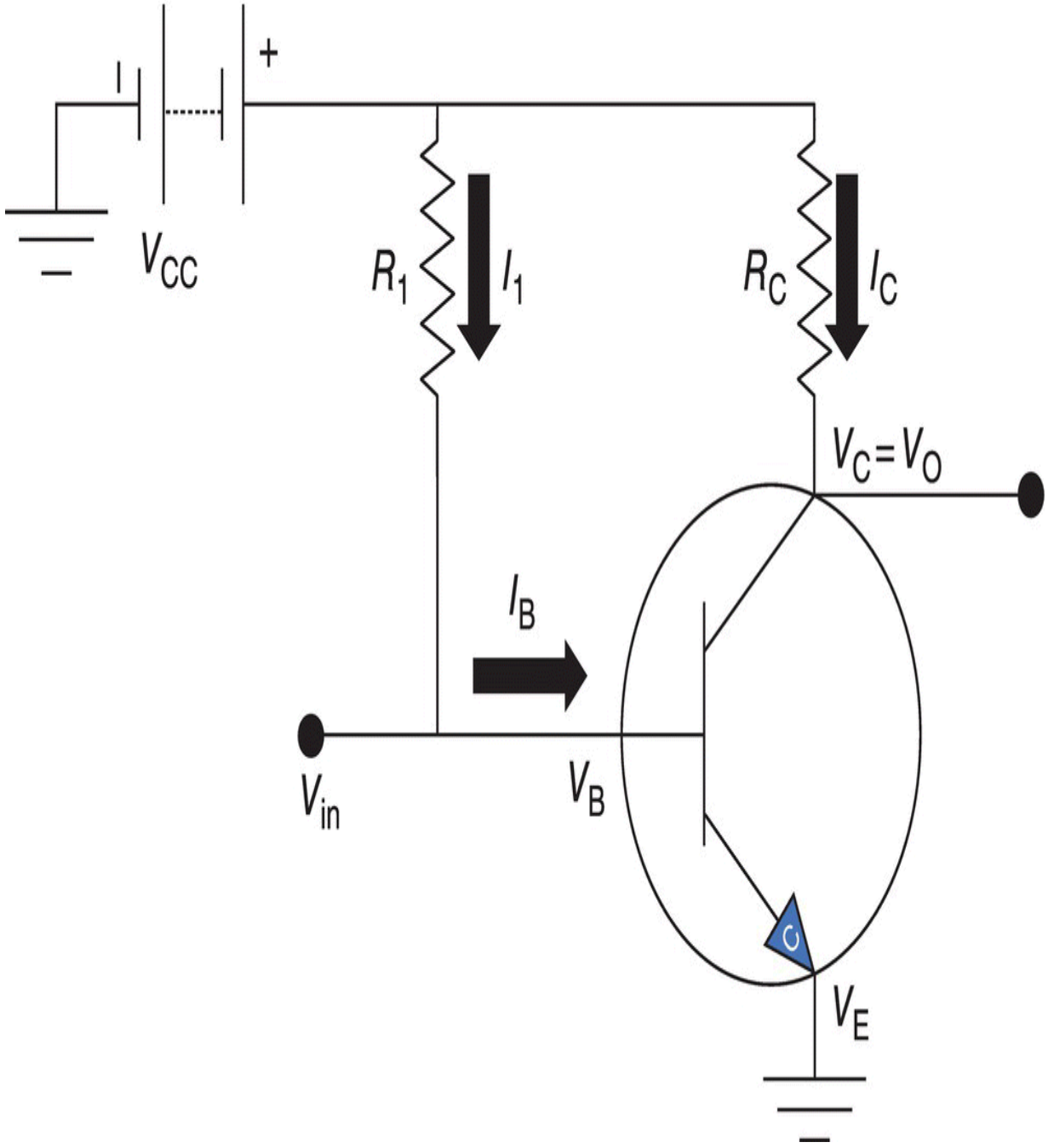


Figure 9.9 The fixed bias circuit is simpler than the collector feedback circuit, but it has less stability against temperature changes.

Notice that V_{CC}, V_{BE} and R_1 do not change, therefore the base current, I_B, is constant no matter what happens to the rest of the

circuit. That is the reason this bias is called the fixed bias circuit. We know that the collector current I_C is

$$I_C = \beta I_B \tag{9.17}$$

and

$$I_C = \frac{V_{CC} - V_{CE}}{R_C} \tag{9.18}$$

I have already pointed out that V_{CC} and R_C are constant, therefore if I_C changes, the only value that can change is V_{CE}.

Let me just point out again that this circuit has no stability. As the temperature goes up, β changes and so does the leakage current, and a positive feedback condition can develop that causes a runaway condition where the current keeps on increasing, especially if the transistor we use has a high β value.

Now let's look again at the transistor characteristic curves in Figure 9.10. There is a slight difference between this load line and the one in Figure 9.4. Yes, when the current I_C is zero the voltage across the transistor is equal to V_{CC}, that has not changed, but now, when the voltage across the transistor is zero, the current must be V_{CC}/R_C. As in the previous case we can draw a load line and select a Q-point.

The calculations are very simple. Looking at the characteristic curve we know that $\beta = 50$, $I_C = 5\,\text{mA}$, and $V_{CC} = 6\,\text{V}$, therefore

$$R_1 = \frac{V_{CC} - V_{VE}}{I_B} = \frac{6 - 0.7}{40 \times 10^{-6}} = 132.5\,\text{k}\Omega \tag{9.19}$$

We'll select a standard resistance value of 130 kΩ.

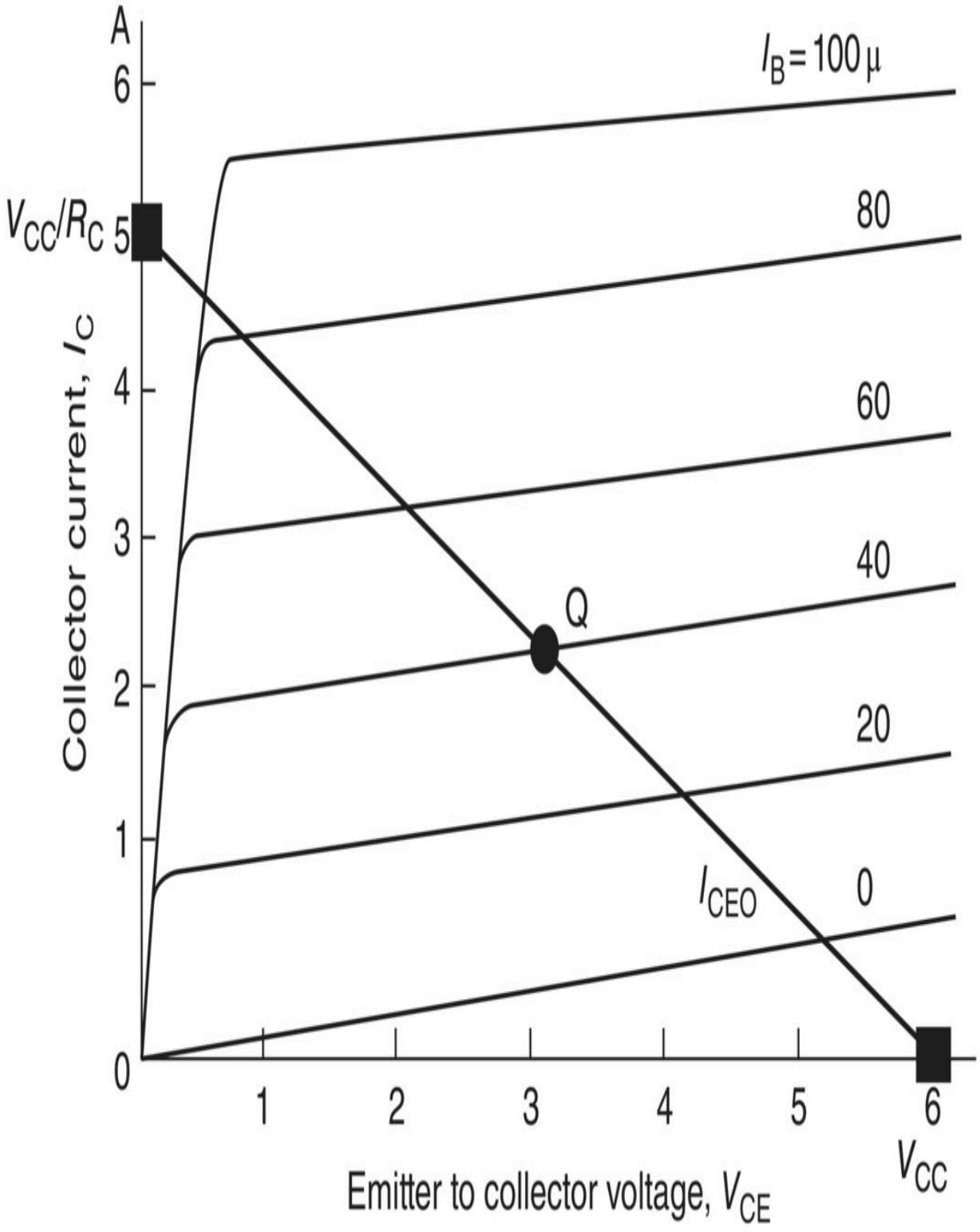


Figure 9.10 The load line on the transistor characteristic curves for a fixed bias circuit is slightly different to the one for the collector feedback circuit.

Now let's take a look at the collector side. At the Q-point since β = 50, then

$$I_C = 50 \times 40 \times 10^{-6} = 2.0 \times 10^{-3} = 2.0\,\text{mA} \tag{9.20}$$

and V_{CE} = 3.2 V, agreeing, obviously, with the Q-values we see in the characteristic curves in Figure 9.10. Therefore, the resistance R_C is

$$R_C = \frac{V_{CC} - V_{CE}}{I_C} = \frac{6 - 3.2}{2.0 \times 10^{-3}} = 1.4\,\text{k}\Omega \tag{9.21}$$

Figure 9.11 is the same as Figure 9.9 but shows the values of the resistors we have calculated.

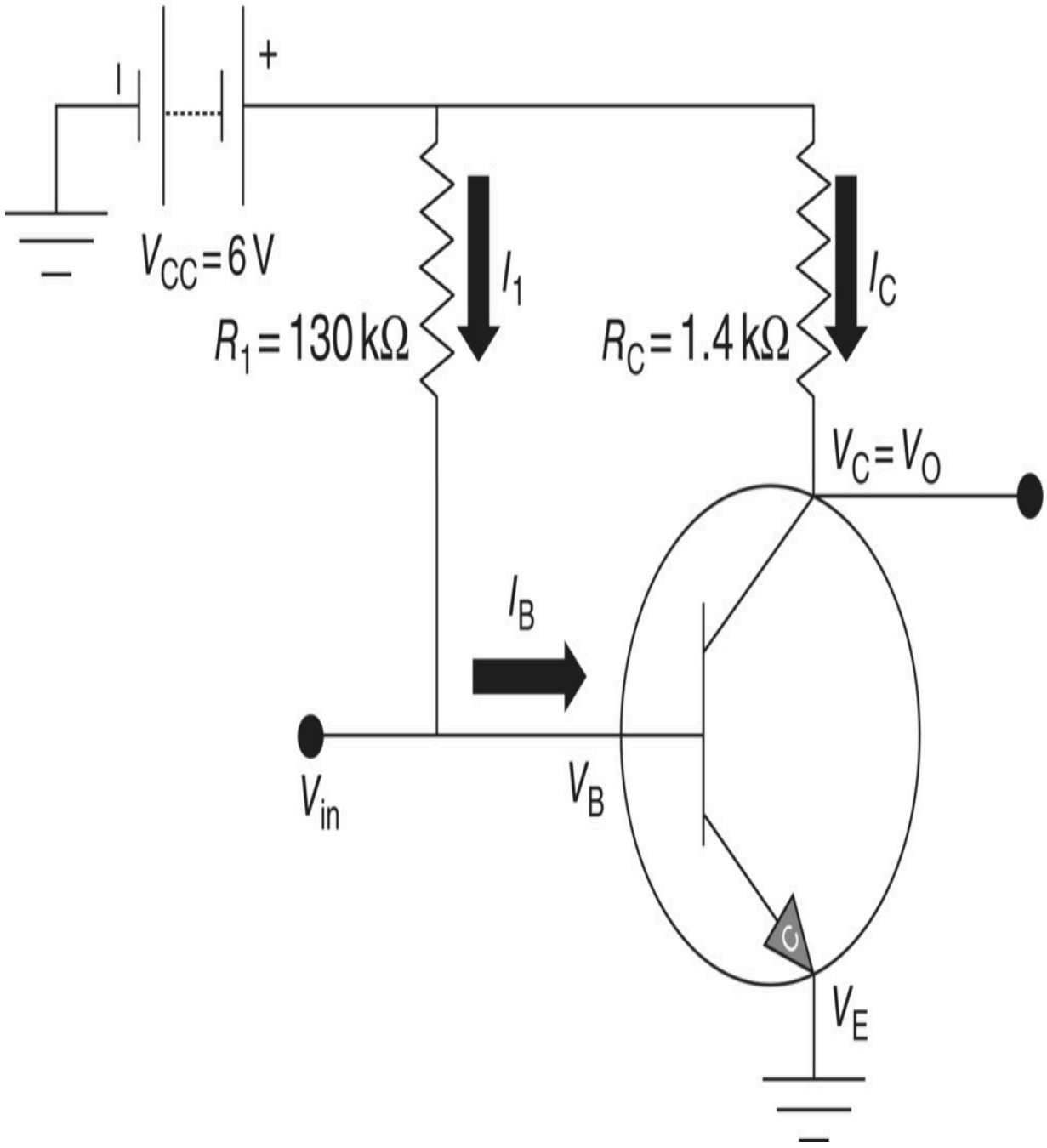


Figure 9.11 The fixed bias circuit with the resistance values we have calculated is not very stable, but it is simpler and can be used in digital circuits where we only care if the transistor is ON or OFF.

9.5 The Collector Feedback Bias Circuit

Just for completeness let me say a few words about the third bias mode, the collector feedback bias mode. I show the circuit in Figure 9.12.

Let see if we can simply show how this circuit stabilizes the transistor operation without going into too many details. First, notice that since I_C is much larger than I_B, by a factor of β, we can say that the current I_C is approximately equal to the total current I or

$$I = I_C + I_B \approx I_C \tag{9.22}$$

Note that the current through the transistor is, approximately, independent of β. Since the emitter is connected to ground, the DC output voltage, V_C, is equal to voltage V_{CE}. As I did with the emitter feedback circuit, let's take a look at how the feedback stabilizes the currents (Figure 9.13).

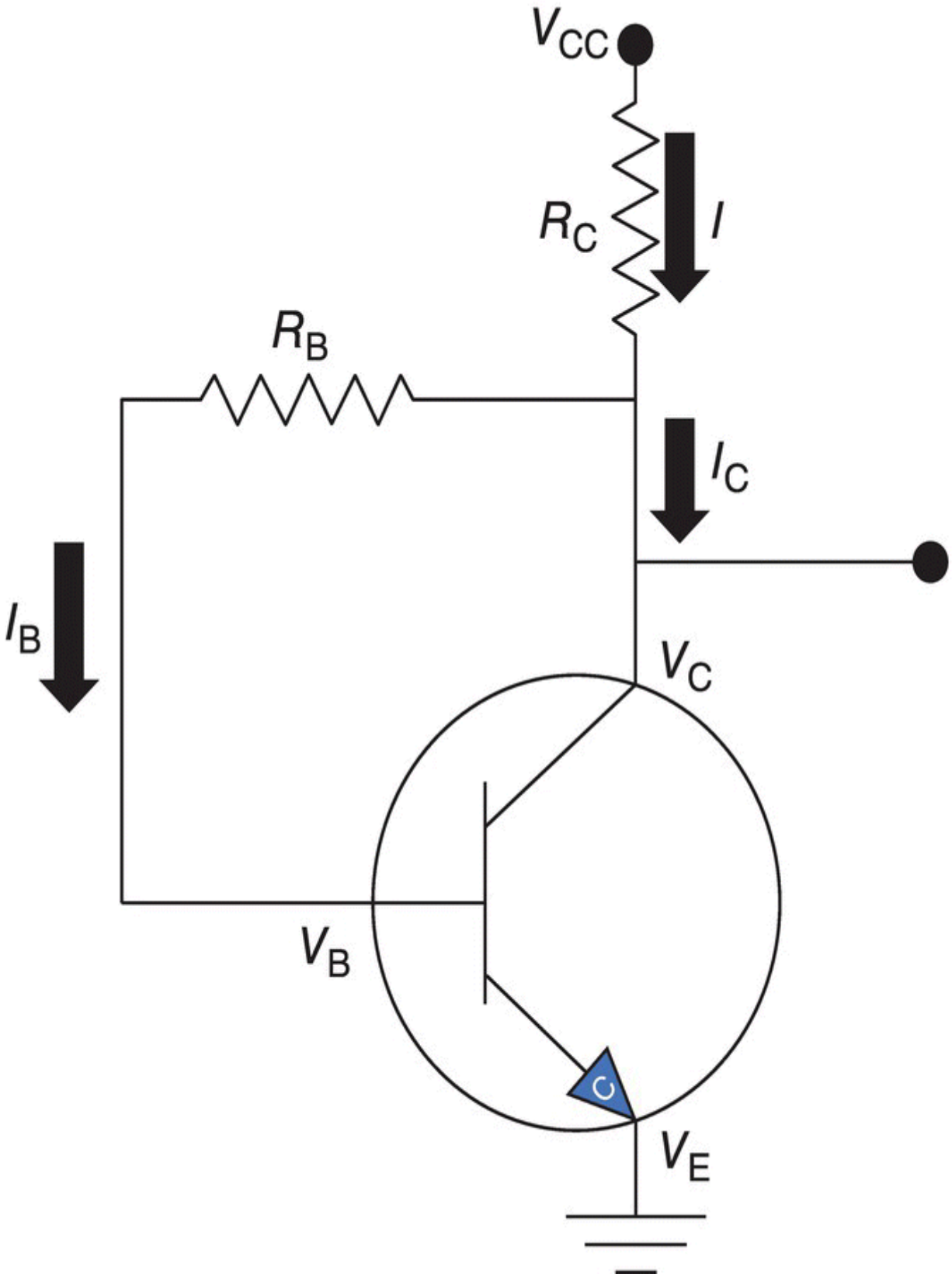

V_{CC}
R_C
I
R_B
I_C
I_B
V_C
V_B
C
V_E

Figure 9.12 The collector feedback bias circuit is a different way of stabilizing the operation of the transistor.

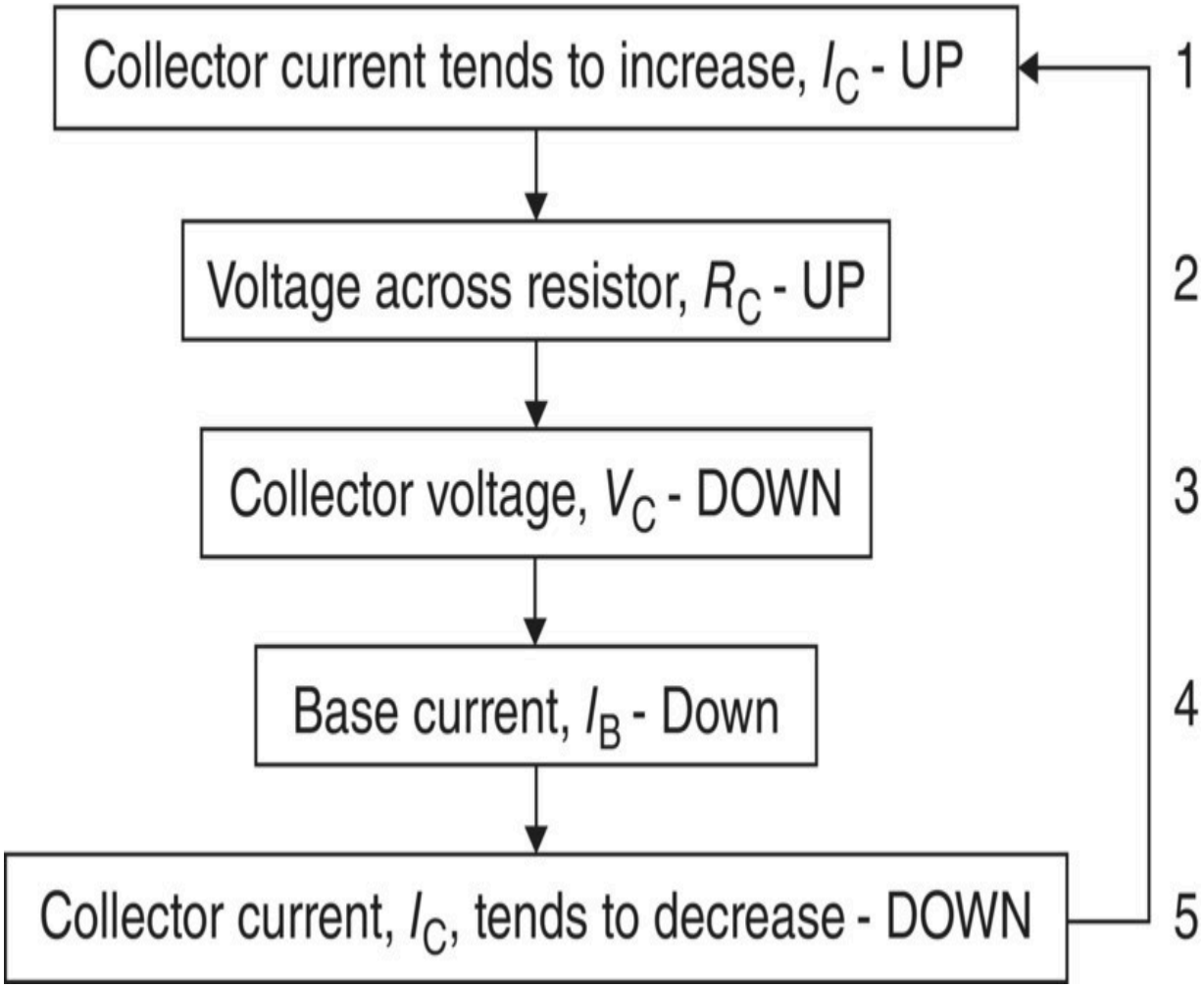


Figure 9.13 Stabilization diagram of the collector feedback circuit.

Following the numbers on the right of Figure 9.13:

Suppose that the collector current I_C tends to go up.

That means that the voltage across the resistor R_C will also tend to go up.

Since the bias voltage V_{CC} is constant, the voltage across the transistor, V_{CE} ($V_{CE} = V_{CC} - V_C$) tends to go down.

The base current, I_B, is equal to the collector voltage, V_C, minus the voltage between the base and the emitter (we already know this is about 0.7 V) divided by the resistance R_B. Since V_C goes down, I_B tends also to go down.

If the current I_B tends to go down, I_C, which is equal to $I_B \times \beta$, tends to go down. This is the opposite of the tendency to go up that we assumed in step 1.

The collector feedback circuit also provides good negative feedback. One of the conditions for this circuit to work, and I expand this in Appendix 9.1, is that β and R_C must be much larger than R_B. Larger R_C or smaller R_B are a problem for the gain of the transistor circuit.

9.6 Power Considerations

Let's talk a moment about power dissipation. I have already mentioned that heat in electronic circuits is a real problem because it not only changes the characteristics of the devices, but also increases the power consumed. So, what is the power dissipation in a transistor? Remember that power is the product of voltage times current or resistance times the square of the current. Thus, the DC power dissipated in the base/emitter circuit is

$$P_{BE} = V_{BE} I_B \tag{9.23}$$

and in the collector to the emitter circuit is

$$P_{CE} = V_{CE} I_C \tag{9.24}$$

Thus, the total DC power dissipated by the transistor is the sum of the two powers. Therefore, in our case, the total power dissipated by the transistor is

$$P_{\text{total}} = P_{\text{BE}} + P_{\text{CE}} \tag{9.25}$$

Notice that when V_{CE} is a maximum, I_{C} is zero (Figure 9.10) and, vice versa, when I_{C} is a maximum, V_{CE} is zero. In these two extremes the power dissipated is zero. The maximum power occurs at the Q-point. In our first case, the emitter feedback bias (Figures 9.3 and 9.4), the maximum power dissipation is

$$\begin{aligned} P_{\text{total}} &= \left(0.7 \times 40 \times 10^{-6}\right) + \left(3.2 \times 2.3 \times 10^{-3}\right) \\ &= 2.8 \times 10^{-5} + 7.4 \times 10^{-3} \approx 7.4\,\text{mW} \end{aligned} \tag{9.26}$$

Notice that the contribution of the base current is tiny compared to that of the collector, so we can disregard its contribution. Now, this is the transistor dissipation only. We also have the dissipation in the resistors. If the transistor is operating at the Q-point, then we can calculate the power dissipated in the resistors in two ways:

$$P_{\text{C}} = R_{\text{C}} I_{\text{C}}^2 = 1.2 \times 10^3 \times \left(2.3 \times 10^{-3}\right)^2 = 6.3\,\text{mW} \tag{9.27}$$

or

$$P_{\text{C}} = I_{\text{C}}\left(V_{\text{CC}} - V_{\text{CE}}\right) = 2.3 \times 10^{-3} \times \left(6 - 3.2\right) = 6.4\,\text{mW} \tag{9.28}$$

The two calculations agree (within my visual approximation of the values in Figure 9.10). Notice that the power of the fixed bias circuit using the transistor I selected is about 14 mW. This may seem very little compared with the 150 W lamp I have in my office but if you consider that a microchip may contain as many 100 million transistors, the chip, if we were to use the transistors we designed, would require a source with multiple kilowatts of power. Obviously, the logic circuits use much smaller transistors that require considerably less voltage and have lower β values.

9.7 Multistage Transistor Amplifiers

Figure 9.14 shows a two-stage voltage amplifier. This is a very trivial way to create a two-stage amplifier. Because the two stages are separated by a capacitor, C_2, the two stages can be biased independently. One transistor may have a higher or lower gain and be biased accordingly. The sinusoidal input signal flows from one transistor to the other until it reaches the output, which I assume, for fun, is a speaker.

Figure 9.15 shows a slight variation on the two-stage amplifier shown in Figure 9.14. The changes are shown inside dotted ovals. One of the problems with capacitor coupled amplifiers is that at very low frequencies the capacitors have a higher resistance (you may recall we call it reactance or impedance, see Appendix 6.1). The impedance decreases the gain as the frequency of the signal gets lower and lower. The capacitor C_6 compensates for that. The resistor R_{C1} is now divided into two resistors, R_{C1a} and R_{C1b}. When the frequency is high, the capacitor C_6 is for all practical purposes shorted, shorting the resistor R_{C1a}. The high-frequency AC signal sees only the smaller resistor R_{C1b}. As the frequency decreases, the capacitance's impedance increases so that the parallel combination of R_{C1a} and C_6 become a larger resistance, and eventually, at zero frequency, the sum of the two resistor equals the original resistor, $R_{C1.}$ If you think about it, this is a very clever idea. The collector resistor/capacitor reactive value keeps on changing automatically as the signal frequency changes.

The second point to note is that you would not design a sound amplifier unless you had a volume control. The addition of the variable resistor, a potentiometer, at the output of the first stage in the amplifier makes volume control possible. If the pointer is set at the bottom of resistor R_P, the voltage input to the second stage is zero, that is, it is grounded. If the pointer is on top of the resistor

R_P, then the maximum signal goes to the second stage. Now we have a reasonably useful sound amplifier.

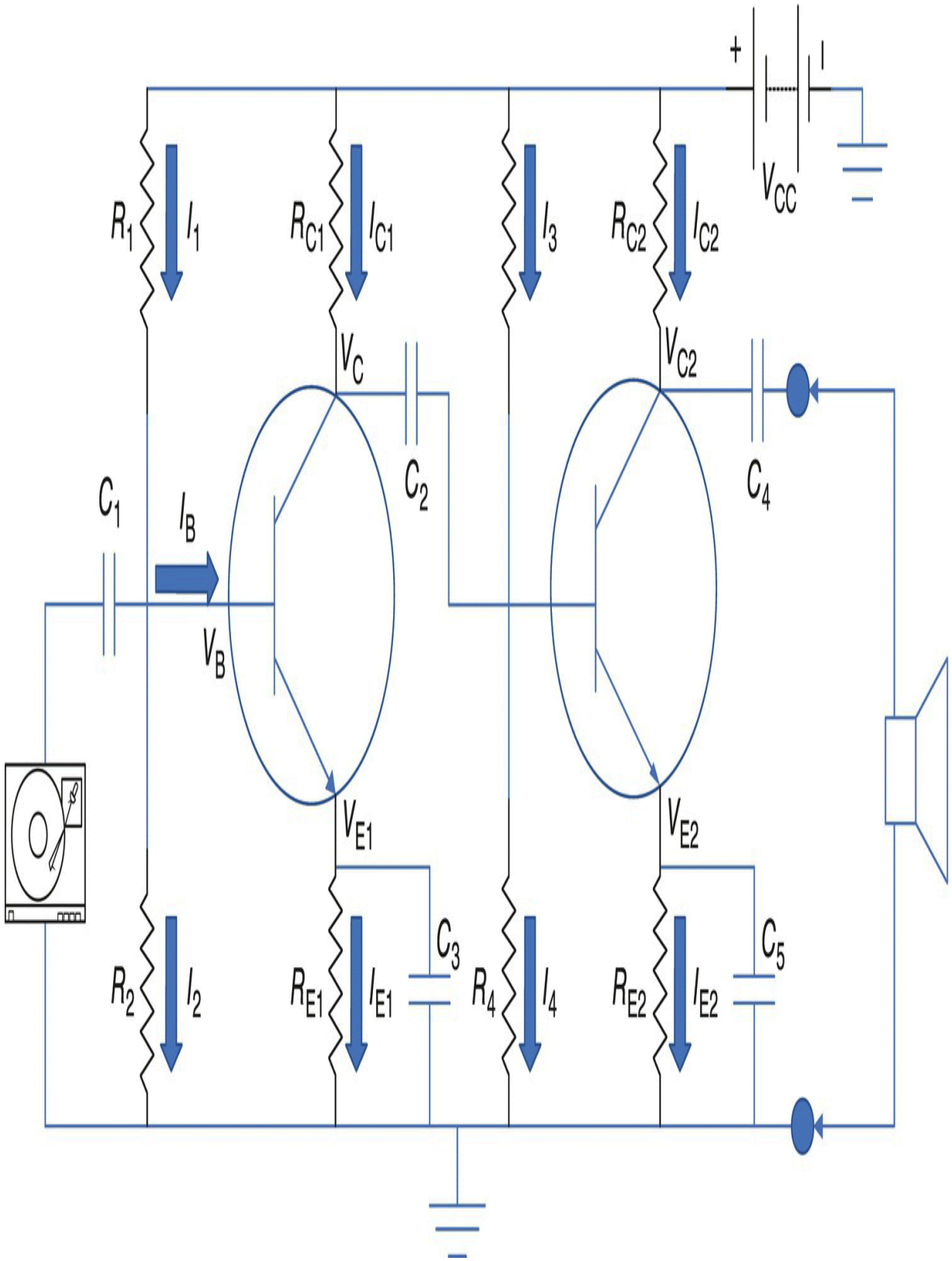

+
V_{CC}
R_1
I_1
R_{C1}
I_{C1}
I_3
R_{C2}
I_{C2}
V_C
V_{C2}
C_1
I_B
C_2
C_4
V_B
V_{E1}
V_{E2}
R_2
I_2
R_{E1}
I_{E1}
C_3
R_4
I_4
R_{E2}
I_{E2}
C_5

Figure 9.14 By connecting two transistor circuits with appropriate capacitors we can amplify sinusoidal signals many times over.

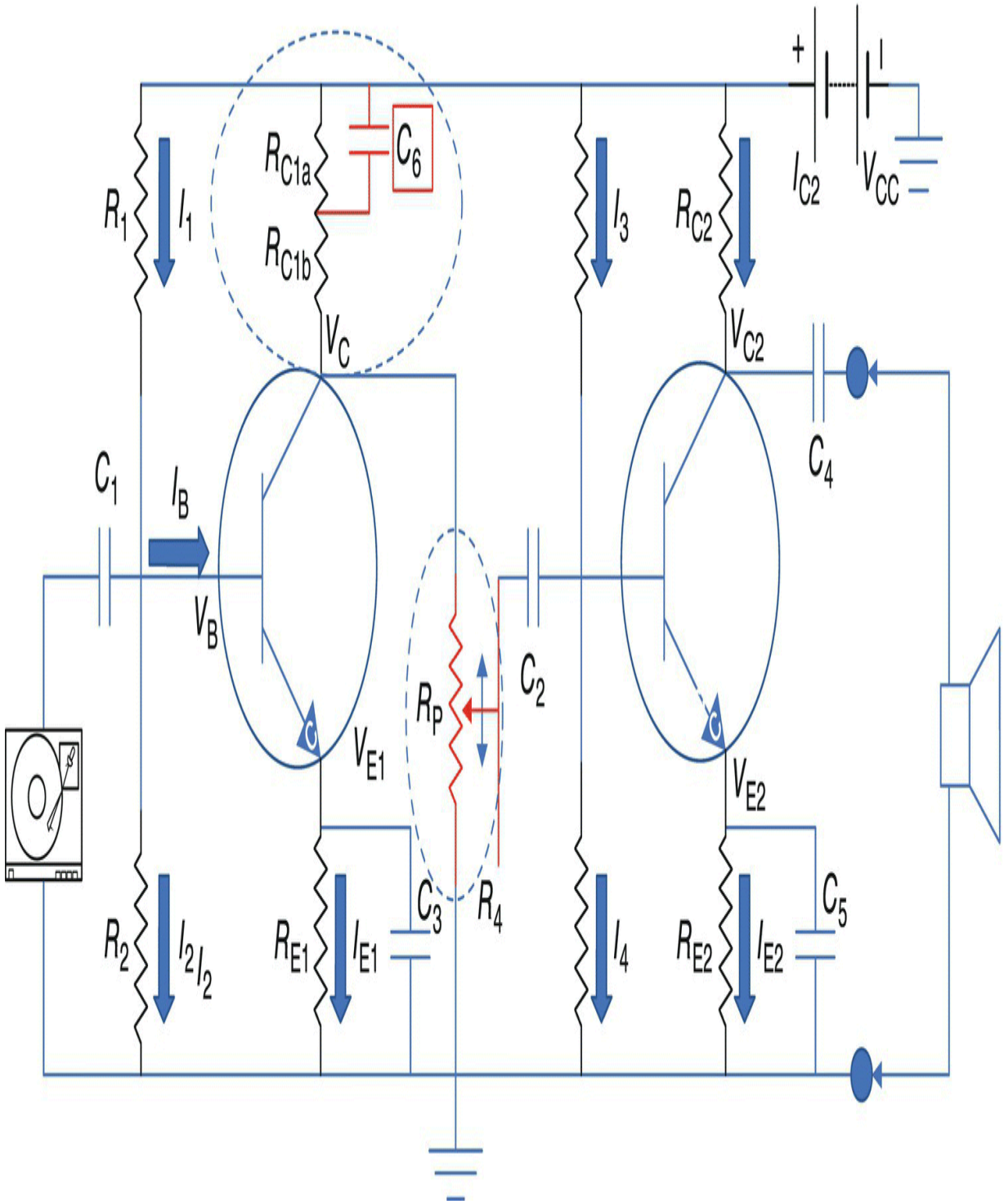


Figure 9.15 By adding a potentiometer and a bypass capacitor (both in red) to a two-stage transistor circuit, we can design a very simple voltage amplifier.

9.8 Operational Amplifiers

In the next chapter I will explain how complex integrated circuits are fabricated, but before I do that, I will talk about operational amplifiers, or OpAmps for short. These are the first implementation of combining many components in a single package. They were developed in the early 1940s and were implemented using vacuum tubes. The OpAmp 741 that I show in Figure 9.16 was developed by Fairchild in 1968 and is still in production. We do not care what is inside, we just use OpAmps in a very large number of applications without worrying about how to bias each individual component. Figure 9.16 shows the inners of a relatively modern OpAmp and Figure 9.17 shows the symbol used for an OpAmp. An engineer designing an electronic device using OpAmps uses the symbol in Figure 9.17 and the characteristics coming from the manufacturer, without caring to understand what is inside the package.

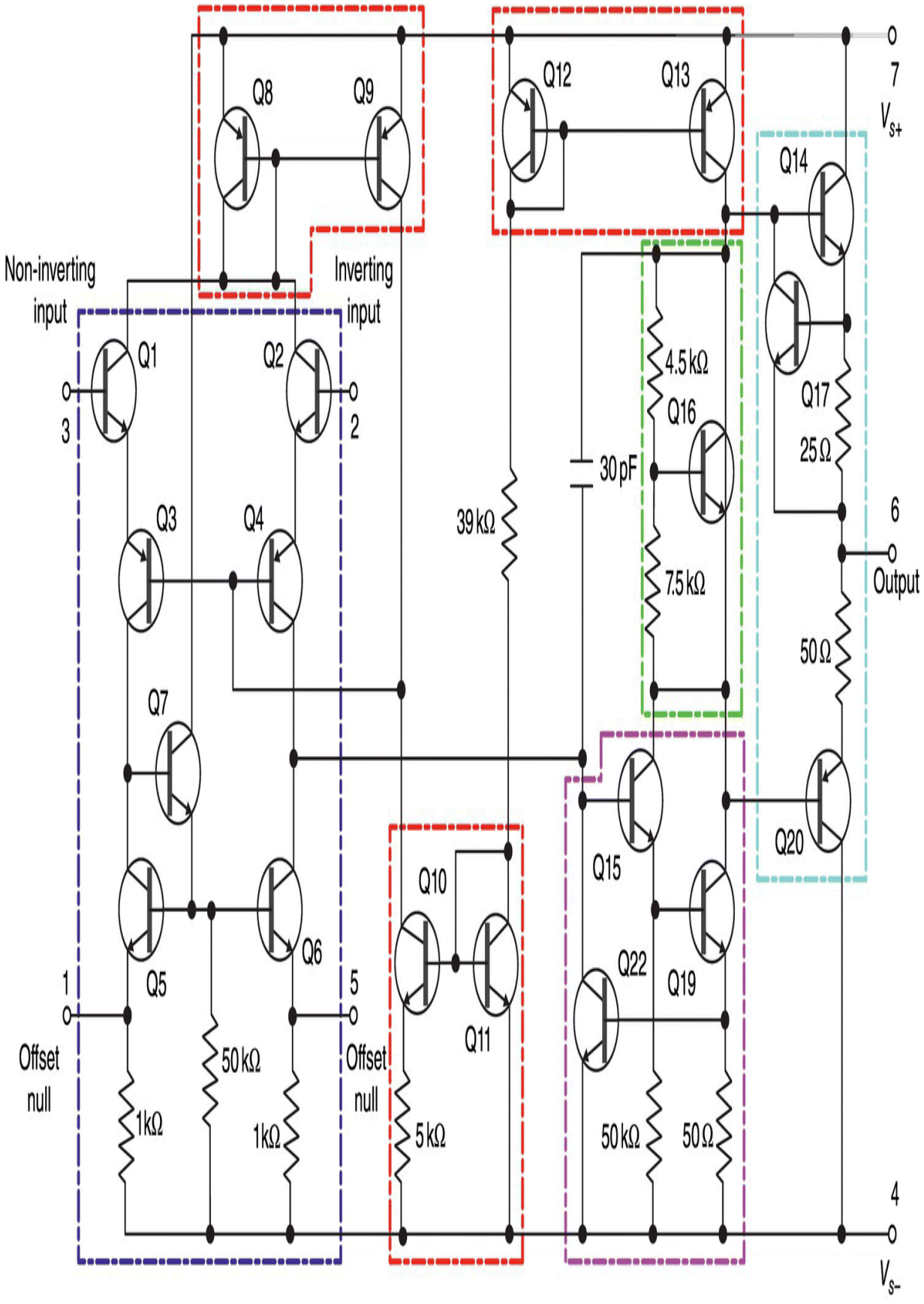

Non-inverting input
Inverting input
Offset null
Offset null
Output
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Q14
Q15
Q16
Q17
Q19
Q20
Q22
1kΩ
50kΩ
1kΩ
5kΩ
39kΩ
30pF
4.5kΩ
7.5kΩ
25Ω
50Ω
50kΩ
50Ω
1
2
3
4
5
6
7
V_{s+}
V_{s-}

Figure 9.16 The internal circuit of an OpAmp, the Fairchild 741.

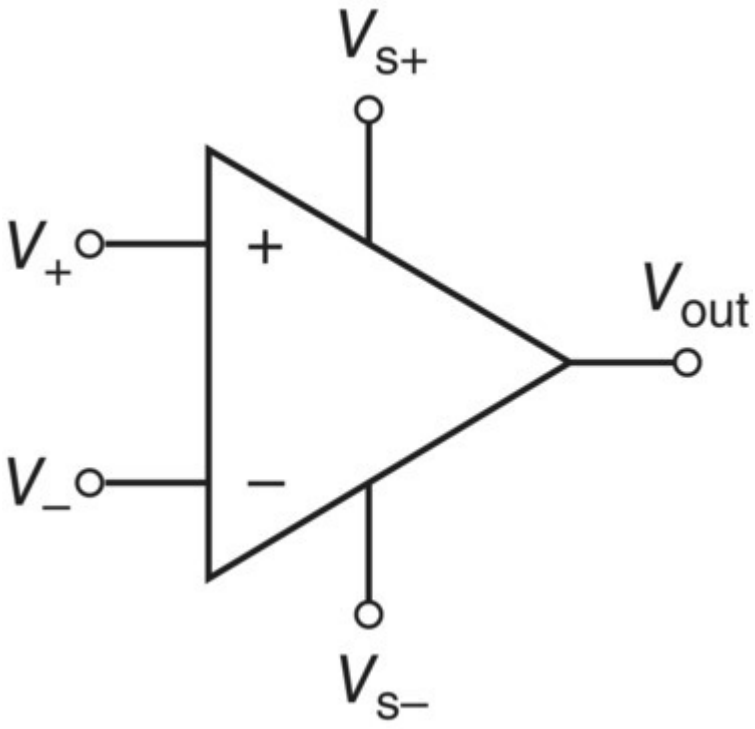


Figure 9.17 The symbol for an OpAmp with two supply voltages, one positive and the other negative, an output, and two signal inputs, one positive and the other negative.

We'll come back to the use of OpAmps, but first I will point out a couple of things about the circuit shown in Figure 9.16. The first is the concept of the differential amplifier. This is the circuit inside the large dotted box on the right of Figure 9.16. Figure 9.18 is a simplification of a differential input circuit. The two transistors in Figure 9.18 are identical (as identical as possible). The two emitters are connected together to a constant current source. These are the two transistors, Q_1 and Q_2, shown in Figure 9.16. The rest of the transistor in that box equalizes the transistor operation and generates a constant current source.

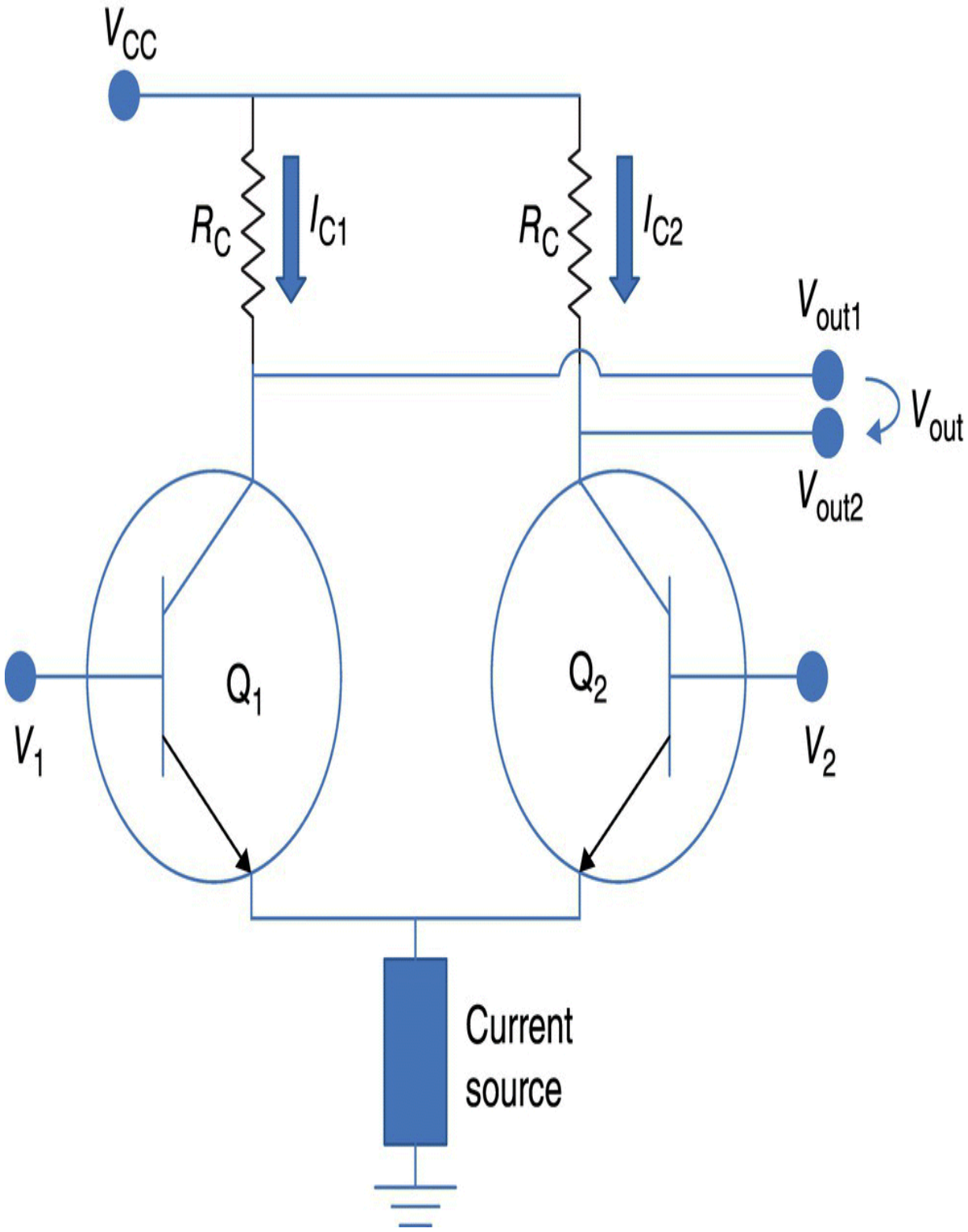


Figure 9.18 A differential input amplifier eliminates many of the noise problems in electronic circuits.

In the circuit of Figure 9.18 we have two inputs, V_1 and V_2, and two outputs, V_{out1} and V_{out2}. We are interested in V_{out}, which is the difference between the two outputs, such that

$$V_{out} = V_{out1} - V_{out2} \tag{9.29}$$

If V_{out1} and V_{out2} are the same, the output voltage, $V_{out,}$ is zero, but if we ground one of the inputs, the output is proportional to whatever the other input voltage is. The main and great advantage of this type of input is that if there is a noise source that interferes with the signal, and we always have electronic noise sources, it affects both inputs, so the output voltage difference is zero. This is a very good way to eliminate the effect of the noisy sources.

Another circuit that is used quite often in an OpAmp is the current mirror. There are three current mirrors in the OpAmp shown in Figure 9.16. Figure 9.19 is a simplified current mirror circuit. In an ideal current mirror, the two transistors, again, are identical. The first thing to notice is that since the two transistors are identical and both emitters are grounded, the two base currents must be the same. There is only one current, I_2, which is split between two identical and equally biased inputs, so

$$I_{B1} = I_{B2} \tag{9.30}$$

Therefore, since I_2 is much smaller than I_{C1}, I can say

$$I_{C2} = \beta I_{B2} = \beta I_{B1} = I_{C1} = I_1 - I_2 \approx I_1 \tag{9.31}$$

Notice all the assumptions I have to make to show that the output and input currents must be the same: the two transistors have to be identical, and β has to be large enough to ensure that I_2 is negligible compared to I_{C1}.

Why do we want a circuit that just replicates the current? The reason is that I am forcing I_{C2} to be the same as I_1 independently of

what circuits are connected to the output.

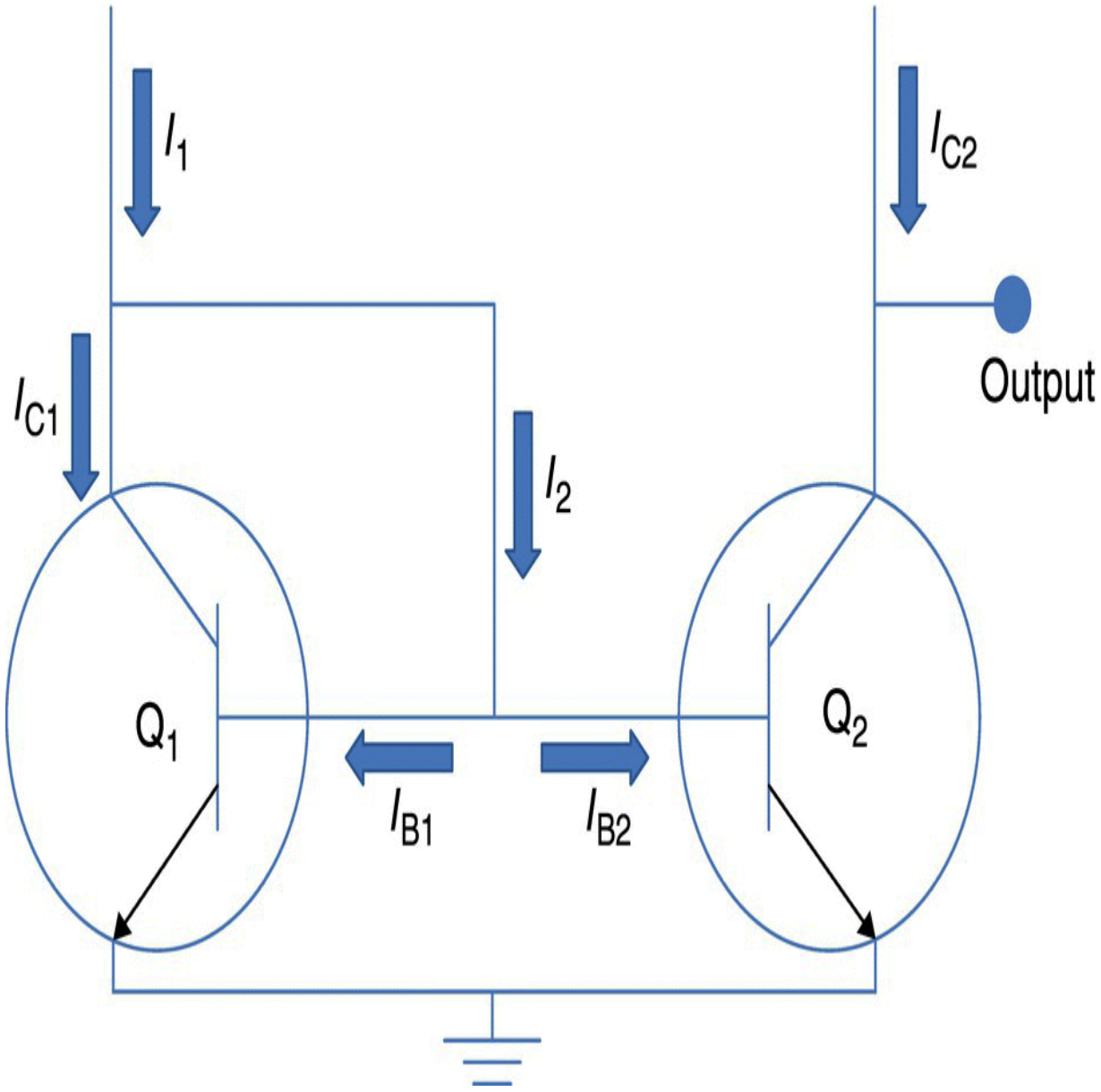


Figure 9.19 A current mirror ensures that the output current, I_{C2}, is the same as the input current, I_I, unaffected by what type of circuit is connected to the output.

9.9 The Ideal OpAmp

Now let's start using OpAmps. Figure 9.20 shows an ideal OpAmp. We assume that the ideal OpAmp has the following characteristics:

Infinite input resistance. This means that the input current is zero, no matter what the input voltage is ($I = V/R$ and something divided by infinity is always equal to zero).

The amplifier has an infinite gain, therefore the input voltage must also be zero (agreeing with point 1) since we cannot have an infinitely high output voltage. Another way to say this is that the input voltage is the output voltage divided by an infinite gain, which, of course, is zero.

The output resistance is zero, so we can have any output current no matter what circuit I connect to the output without losing any voltage.

Now let's consider ways of using this ideal OpAmp. Figure 9.21 shows a typical inverting circuit. Let's see what this circuit does. Since the positive input, V_+, is connected to ground, 0 V, and there cannot be a voltage between the two input terminals (remember the base current has to be zero), the negative input, V_-, also has to be 0 V. So, the voltage at point A in Figure 9.21 is zero. Furthermore, since I_B is also zero, the current I_2 must be equal to the current I_1, therefore (remember point A is zero)

$$I_1 = \frac{V_{in}}{R_1} = I_2 = \frac{V_{out}}{R_2} \tag{9.32}$$

If we now calculate the voltage gain, we find a very simple relationship

$$\text{voltage gain} = \frac{V_{out}}{V_{in}} = -\frac{R_2}{R_1} \tag{9.33}$$

Notice the beauty of this "ideal" OpAmp. I can get any voltage gain I want just by choosing the value of two resistors. You want a gain of 1? Make $R_2 = R_1$. Do you want a gain of 10 million? Make $R_2 = 10\,000\,000 \times R_1$. (Of course, I'm kidding. You know things are not

ideal.) The actual characteristics of OpAmp 741, one of the most commonly used OpAmps, are:

input resistance $>0.1\ \mathrm{M\Omega}$

output resistance $<100\ \Omega$

voltage gain $\approx 2\,000\,000$.

Characteristics of the ideal OpAmp

R_{in}	Infinity
R_{out}	Zero
I_b	Zero
Gain	Infinity

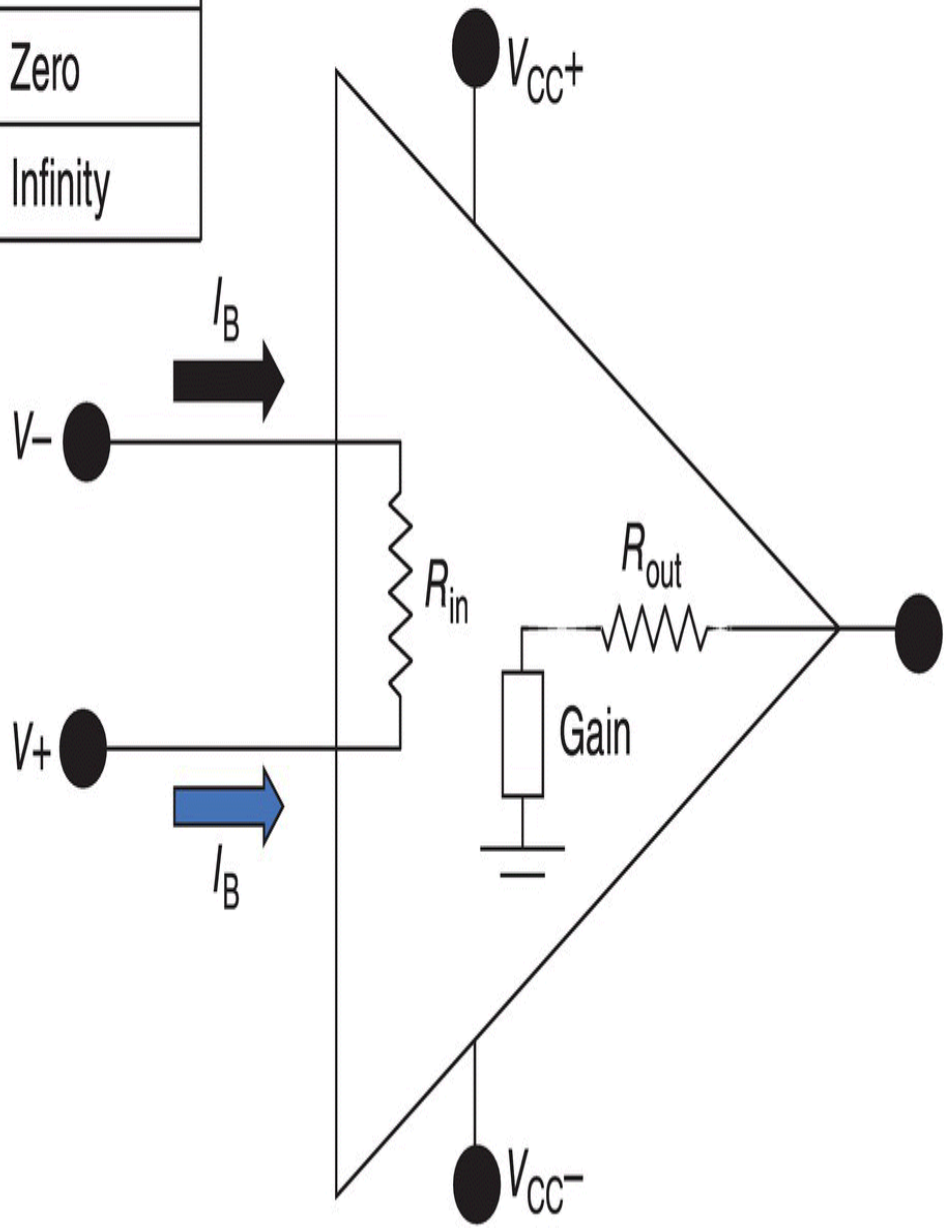


Figure 9.20 The ideal OpAmp has an infinite resistance, zero output resistance, and an infinite gain.

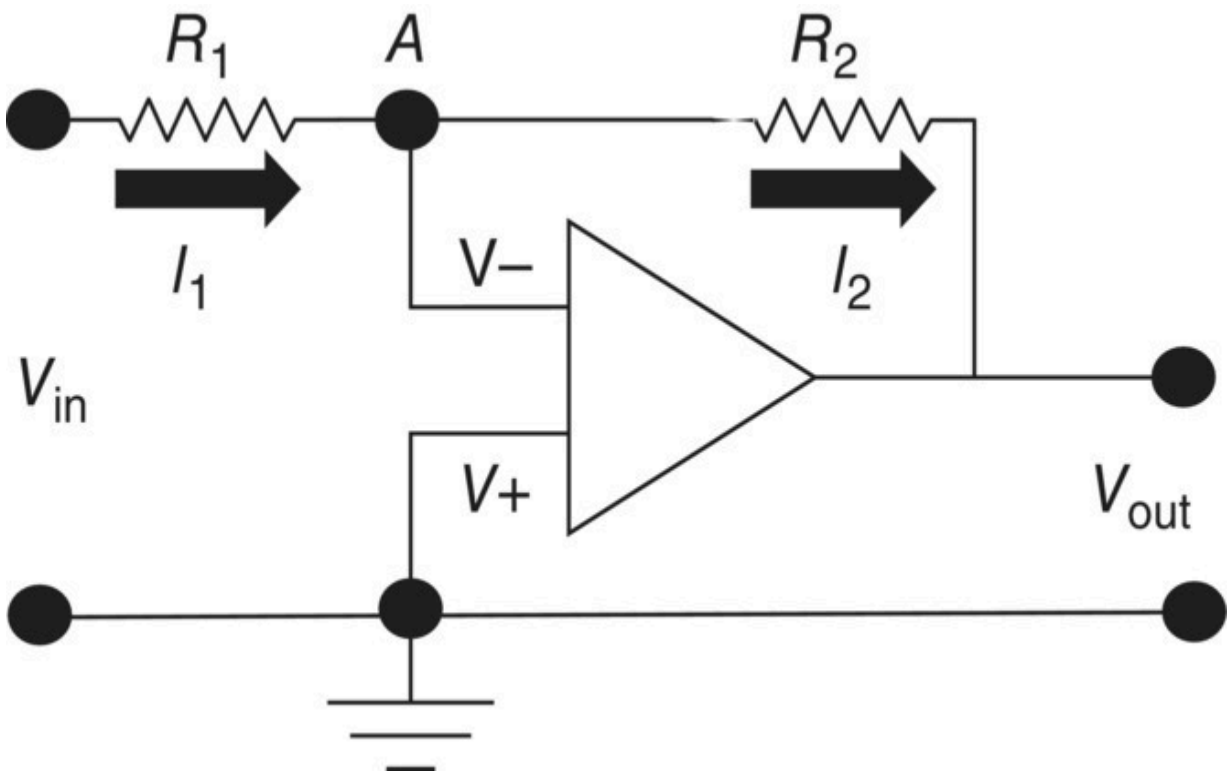


Figure 9.21 An inverting OpAmp has a gain defined by the ratio of the two resistors, $-R/R_1$.

I will mention one further application, the differential amplifier, which I show in Figure 9.22. In this differential circuit we have two different input voltages, V_1 and V_2. V_1, R_1, and R_2 are connected to V_- and form the same inverting circuit that I show in Figure 9.21. Now, instead of grounding V_+, I connect it to a different voltage V_2 with a voltage divider provided by resistors R_3 and R_4. From what we learned about the inverting circuit (Figure 9.21) we know that

$$I_1 = \frac{V_1 - V_-}{R_1} = \frac{V_- - V_{out}}{R_2} \quad \text{and} \quad I_2 = \frac{V_2 - V_+}{R_3} \tag{9.34}$$

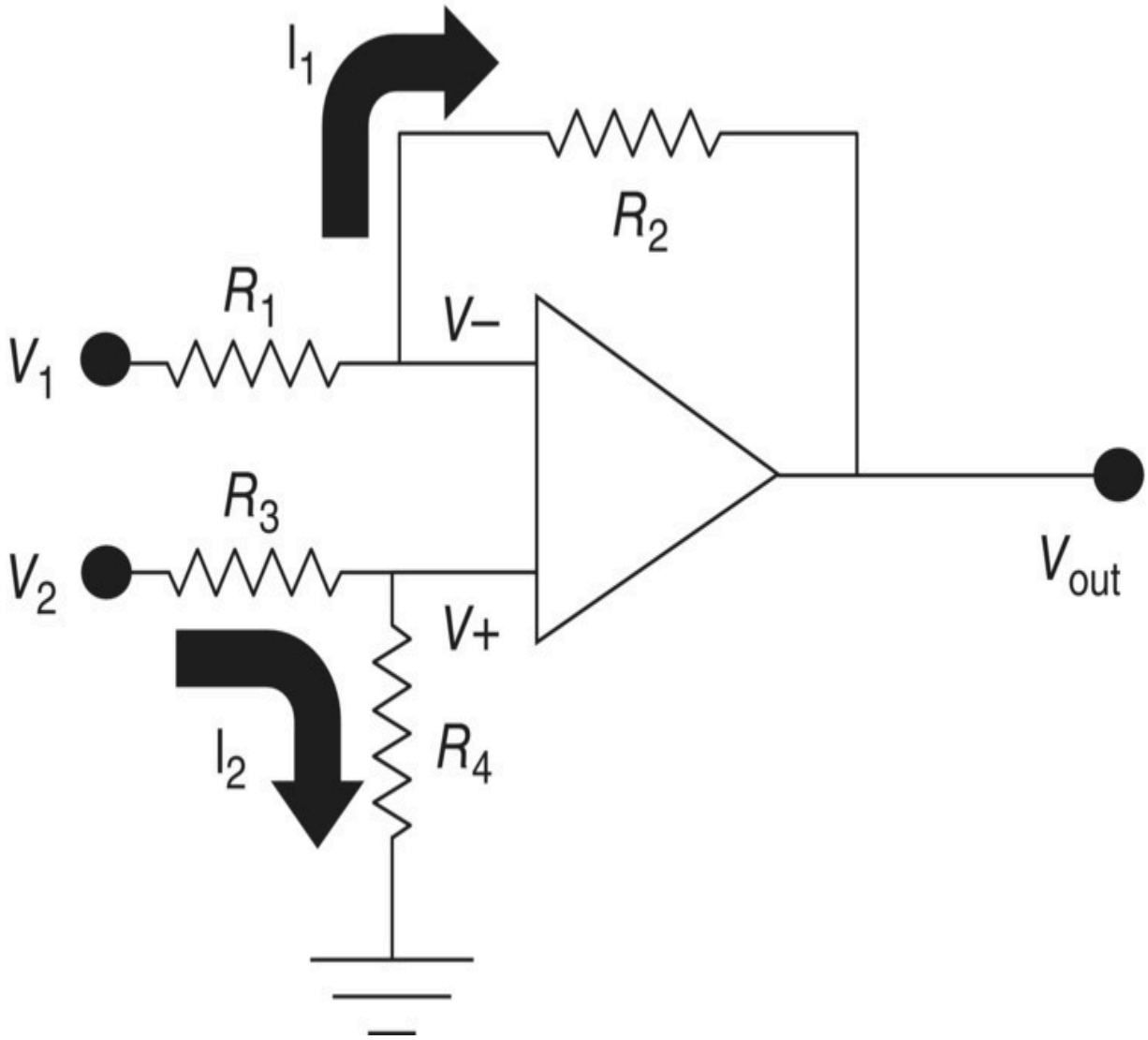


Figure 9.22 A differential amplifier provides a gain defined by the ratio of two resistors and is free of noise.

We can also say that, because of the voltage divider,

$$V_+ = V_2 \frac{R_4}{R_3 + R_4} \tag{9.35}$$

Now the output V_{out} is going to be the contribution of the two inputs, or

$$V_{out1} = -V_1 \frac{R_2}{R_1} \quad \text{and} \quad V_{out2} = V_2 \left(\frac{R_4}{R_3 + R_4} \right) \left(\frac{R_1 + R_2}{R_1} \right) \tag{9.36}$$

If I make all the resistances the same, I get

$$V_{out} = V_{out2} + V_{out1} = V_2 - V_1 \tag{9.37}$$

which is a unit gain differential amplifier, gain of 1, but any noise that has been picked up will cancel out. If we desire some gain, we make R_1 equal to R_3. and R_2 equal to R_4. Then

$$V_{out} = -V_1 \frac{R_2}{R_1} + V_2 \left(\frac{R_2}{R_1 + R_2} \right) \left(\frac{R_1 + R_2}{R_1} \right) = \left(V_2 - V_1 \right) \frac{R_2}{R_1} \tag{9.38}$$

which gives the same gain as we obtained in the inverter OpAmp except that now we have the differential operation, which cancels the noise. Notice that if I make V_1 equal to zero, I get the same gain relation as I show in Eq. (9.33).

9.10 Summary and Conclusions

In this chapter we learned how to bias a transistor to get some useful devices. Because key transistor parameters like leakage current and current gains change with temperature, I discussed three ways of biasing the transistors, some with negative feedback that cancels the effects of the changes by having opposite trends forcing the gain to go down when the collector current goes up and vice versa. We saw how the characteristic curves allow us to choose the operating values, voltages, and currents so the transistors operate in the linear region and the signals are amplified without distortion.

By isolating or shorting portions of the circuit using capacitors we have seen how sinusoidal voltages are amplified without disturbing the DC conditions that make the transistor functional and stable.

We have seen also the concept of how transistors circuit can be added and combined with others to fabricate useful amplifiers. In the process I discussed the first and primitive type of integrated circuits, OpAmps, that make useful electronic devices much easier to design and fabricate.

In this chapter we have used a lot of arithmetic to explain the behavior of the basic circuit. In next chapter we'll relax and examine how these integrated circuits are fabricated. No more arithmetic for a while.

Appendix 9.1 Derivation of the Stability of the Collector Feedback Circuit

In this appendix I want to more formally demonstrate why the emitter feedback bias is quite stable. Take a look at Figure 9.23.

The total current, I, is divided between the collector and the base currents, and since I know that $I_C = \beta I_B$, then I can write

$$I = I_C + I_B = \beta I_B + I_B = (\beta + 1) I_B \tag{9.39}$$

Following the outside loop, the total voltage, V_{CC}, must equal the sum of the voltage drop across the two resistors plus the voltage between the base and the emitter, which we know in a transistor is about 0.7 V. So,

$$V_{CC} = IR_C + I_B R_B + V_{BE} = \left[\left(\beta + 1 \right) R_C + R_B \right] I_B + V_{BE} \qquad (9.40)$$

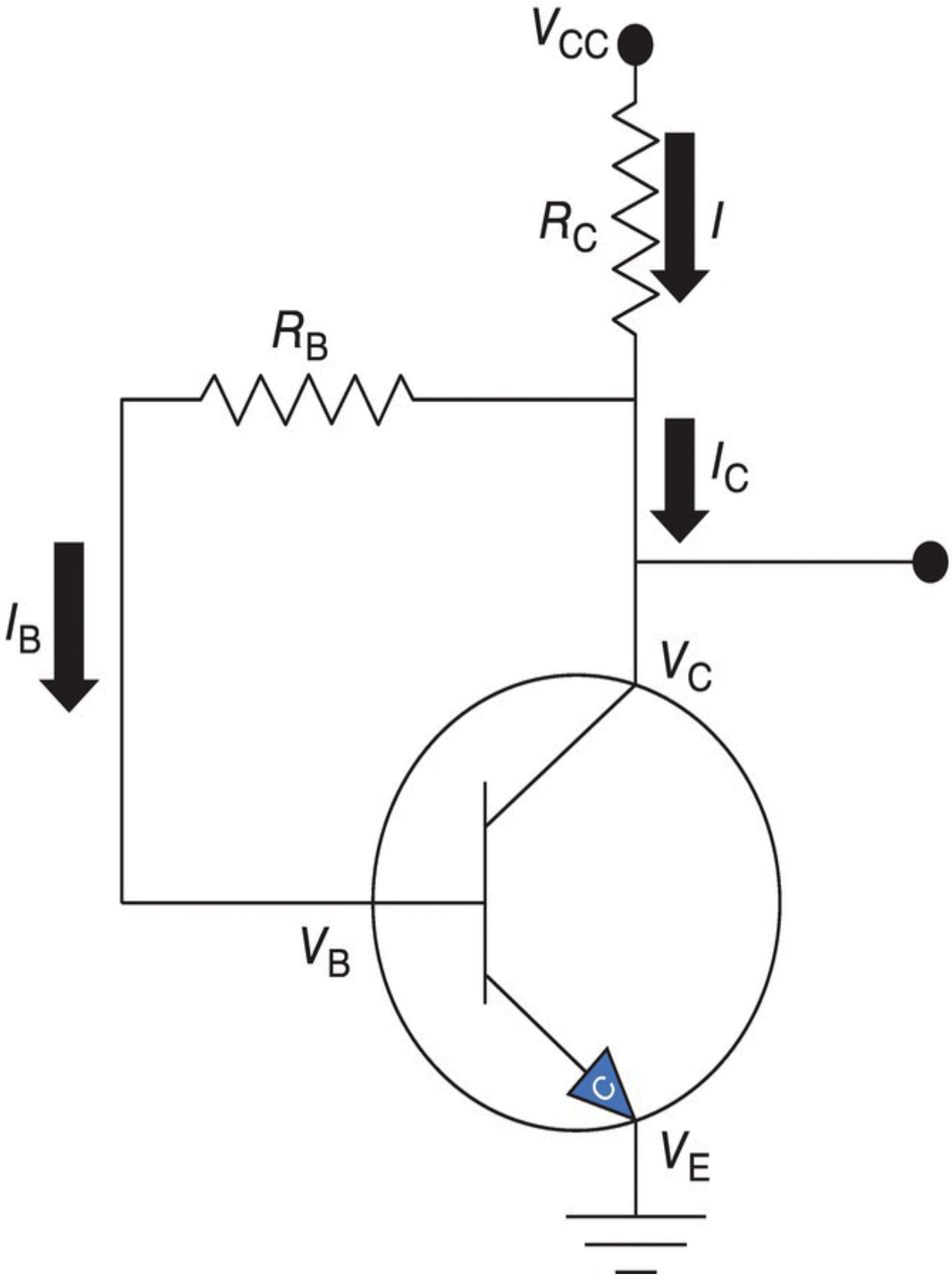

V_{CC}
R_C
I
R_B
I_C
I_B
V_C
V_B
C
V_E

Figure 9.23 The collector feedback bias circuit is a different way of stabilizing the operation of the transistor (same as Figure 9.12, repeated here for convenience).

I can solve equation (9.40) for the base current, I_B:

$$I_B = \frac{V_{CC} - V_{BE}}{(\beta + 1)R_C + R_B} \tag{9.41}$$

The collector current, I_C is

$$I_C = \beta I_B = \frac{\beta(V_{CC} - V_{BE})}{(\beta + 1)R_C + R_B} \tag{9.42}$$

Now I know that β is much larger than 1, so dropping the 1 in the denominator of Eq. (9.42) I get

$$I_C \approx \frac{\beta(V_{CC} - V_{BE})}{\beta R_C + R_B} \tag{9.43}$$

Now if I make $\beta R_C >> R_B$, then I can drop R_B and the βs cancel out and we end up with

$$I_C \approx \frac{V_{CC} - V_{CE}}{R_C} \tag{9.44}$$

Therefore, I_C is independent, approximately, from β. Again, approximately, β can change but the collector current will not change much, and we get the stabilization we desire.

I have made a point of using the word "approximately" twice. That is the problem with the collector feedback circuit versus the first one we discussed, the emitter feedback bias. For collector feedback bias to work it requires $\beta \times R_C$ to be much larger than R_B. Since often we do not know the value of β as it changes, we need to make R_C

larger than R_B, which implies that V_{CC} has to be larger, increasing the cost and the power dissipation, or R_B must be smaller, decreasing the reversed bias value between the collector and the base junction.

As is true in any design, the engineer has to balance the advantages and disadvantage of any design. Nothing is truly free.

10
Integrated Circuit Fabrication

OBJECTIVES OF THIS CHAPTER

We now understand the physics and some of the uses of the most common semiconductor electronic components. It is time to see how we make them. The methodology of how semiconductor devices are made is quite simple and is basically the same as was used 60 years ago. Sure, the technology has improved by factor of millions and the equipment used is so sophisticated that one piece of equipment can cost a billion dollars, but what these sophisticated machines do is quite easy to understand.

In this chapter we are going to look at all the processing steps that we use to make semiconductor devices. You will see that the process is a repetition of a number of steps. As I discuss different processes, I will exemplify them by showing step by step how we fabricate a simple transistor (Sections 10.5 and 10.7). Believe me, the circuits can be very complicated but the steps to get there are the same (almost) as those used to fabricate a single transistor. I recommend you go to the inserts with the color figures to better understand the process.

10.1 The Basic Material

The basic material to obtain electronic quality silicon is sand or silicon dioxide. With a variety of chemical reactions and high-temperature treatments, converting silicon first into a high-purity gas and then solidifying again, we are able to remove many of the impurities (primarily iron and aluminum). Because these impurities

or impurity compounds have different boiling points, we are able to get silicon with a purity better than one impurity per 5×10^{13} cm^{-3} or about 99.9999999% pure. This is quite a pristine and pure silicon, but at this point the silicon crystals are arranged randomly, what we call polysilicon, a kind of an amorphous material, like a number of small randomly arranged mosaics. We need large areas of perfectly arranged silicon atoms in the form of a single crystal. We also want the crystals to grow in a particular orientation.

10.2 The Boule

What we want is a long, cylindrical bar of as pure a silicon, free of defects and impurities, as we can get and with a perfect crystallographic lattice. That's all we want! We call it a silicon boule. There are basically two techniques we use to get this almost perfect boule: the Czochralski method and the float-zone method.

10.2.1 The Czochralski Method

The main technique to manufacture a boule of silicon is the Czochralski method. Jan Czochralski (1885–1953), a polish scientist, Figure 10.1, invented this method in 1915, long, long before we thought about electronic circuits.

The Czochralski process is the favorite way to grow pure or controlled doped silicon boules. I show the process in Figure 10.2. We put into a crucible chunks of the purest polysilicon material we can get, and heat it at a temperature above the melting point of silicon, usually above 1500 °C (the melting point of silicon is 1412 °C). We then dip a small silicon crystal, that we call the seed, of the right properties and crystallographic orientation, into the silicon melt and start pulling it up very slowly at the same time that we rotate both the seed and the crucible, in opposite directions. The speed of rotation and the vertical pull determine the diameter of the final boule. As the silicon is pulled out of the melt, the surface tension holds the melt together, which solidifies and attains the same

crystallographic properties as the seed, the same as the yeast in baking bread. The slower it goes up, the wider the cylindrical silicon rod, the boule, is going to be.

Figure 10.1 Dr. Jan Czochralski developed a method of growing very pure and uniform crystals, called boules.

Source: https://upload.wikimedia.org/wikipedia/en/9/99/Jan-czochralski.jpg.

Often we want the boule to be n- or p-type. In this case we add the appropriate gasses, such as boron for p-type or phosphorous for n-type, as we pull the boule up.

Currently, we can fabricate boules that are 300 mm in diameter, close to a foot, and there is research going on to fabricate a 450-mm boule, a whopping one and a half feet in diameter. The larger the diameter the more chips we can fabricate in a wafer and the cheaper each chip is going to be. In the 1960s and 1970s we were happy with 2- and 4-in. diameter boules. It may take up to three days to get one of these much larger boules out of the melt.

Once the boule is grown, we grind the uneven surface to form a perfect cylinder. The subsequent processes require that the wafers are all exactly the same size. We make a very small notch or flat cut on one side of the boule to identify the crystallographic orientation of the wafers after they are cut. I explain the reasons for this process in Appendix 10.1.

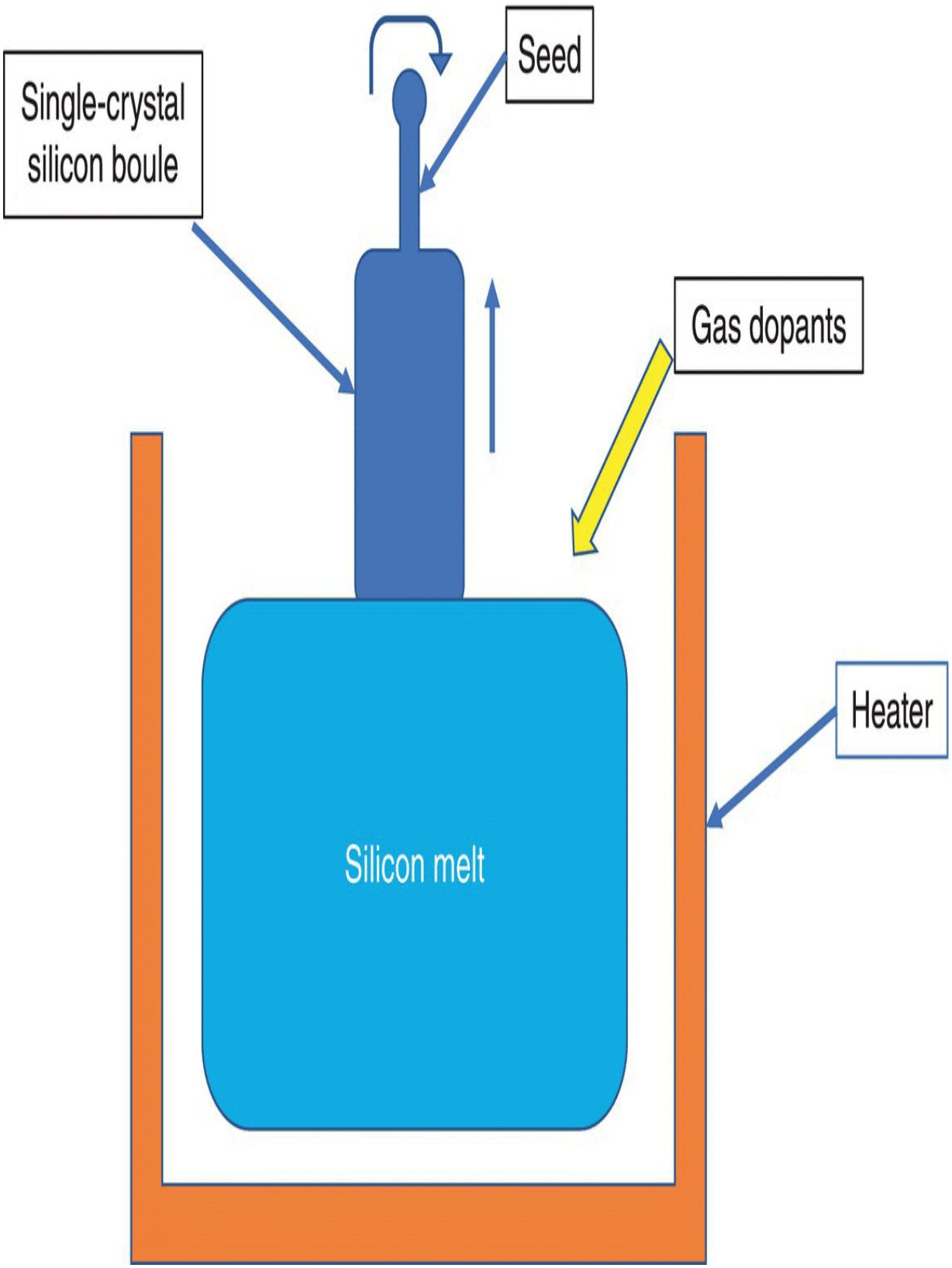

Seed
Single-crystal silicon boule
Gas dopants
Heater
Silicon melt

Figure 10.2 The Czochralski method to grow a silicon boule. A seed pulls the boule from the melt, while rotating it slowly as it moves up.

10.2.2 The Flow-zone Method

I show the float-zone method in Figure 10.3. We place a boule of polysilicon (dark blue), as pure as we can get, in a vacuum chamber. A polysilicon boule is just a rod of silicon in which atoms are arranged in a random way. As in the Czochralski process, there is a crystalline silicon seed (light blue) at the bottom. A high-temperature radio-frequency heater (red dots) moves slowly up. As the heater moves up, it melts the polysilicon (orange region) and then the melt solidifies in an orderly crystalline boule (light blue). Furthermore, we sweep many impurities up due to the fact that impurities have different segregation coefficients. Luckily for us, the impurities prefer to stay in the melted region rather than move into the solid crystalline boule. So, when we finish the process, we just cut the two ends of the boule where the impurities have concentrated. The process can be repeated, going up and down a few times, to obtain higher and higher purity boules. The surface tension prevents the two solid regions of the boule, the polysilicon (dark blue) and the crystalline (light blue), from falling apart but, because of the weight, this limits the size of the boule. The fact that the entire boule is in a vacuum chamber results in very low oxygen content. Oxygen provides strength to the wafers so that they can withstand the many heating steps required for integrated circuit fabrication. That is another reason why the Czochralski method is preferred.

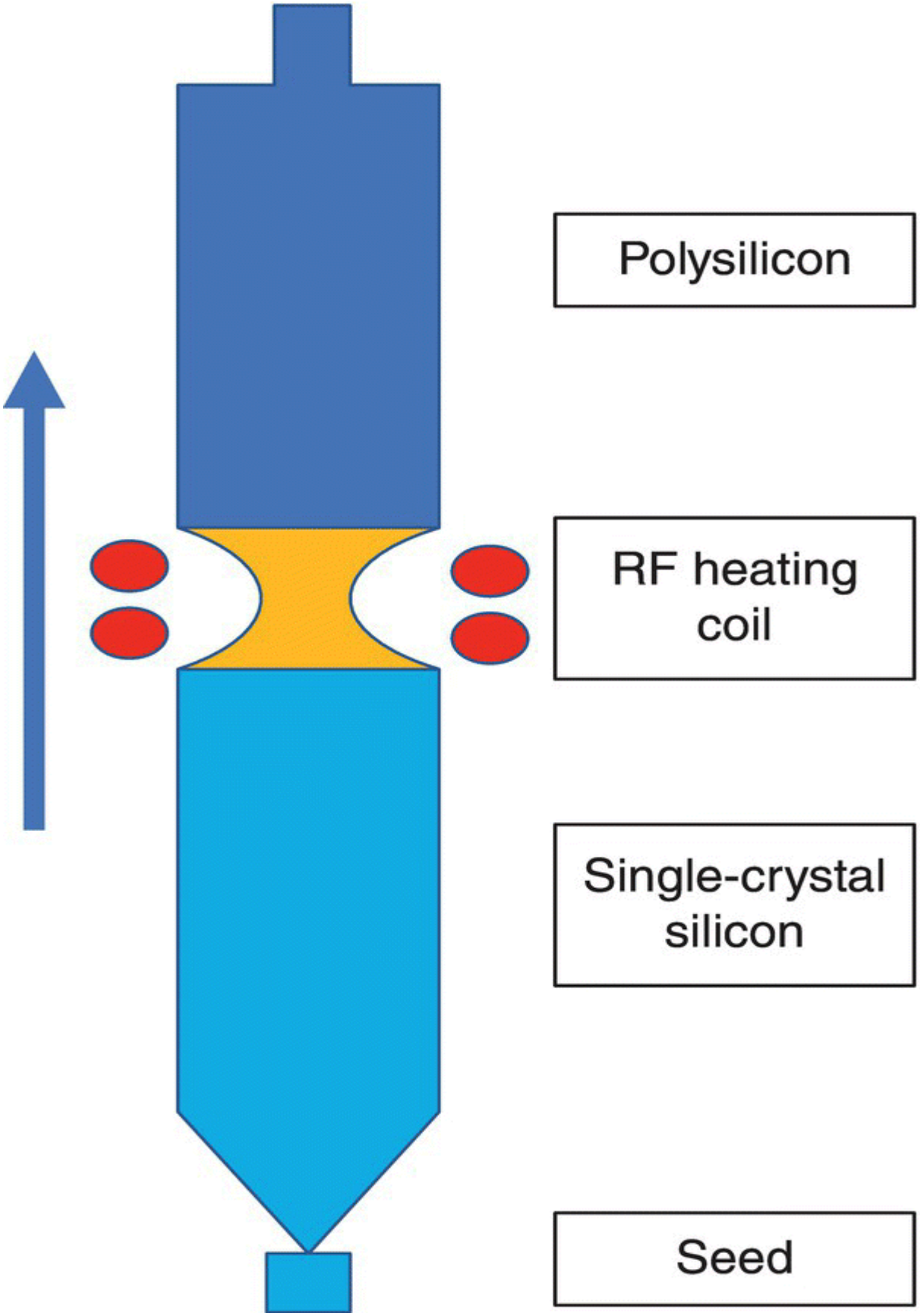

Polysilicon
RF heating coil
Single-crystal silicon
Seed

Figure 10.3 In the float-zone growth method a heating coil moves up and down, melting and recrystallizing the boule and segregating impurities to the two ends of the boule.

10.3 Wafers and Epitaxial Growth

Once the boule is completed, we slice it into very thin wafers, about 1/32 of an inch thick. We want the wafers to be very flat so that when we project the optical patterns onto its surface the entire wafer is in focus. The wafers are polished both mechanically (abrasive polishing) and chemically, then they are cleaned and inspected, ready to enter the process.

In most cases we want a more perfect material than the boules can provide and we want to introduce different dopings. We grow epitaxial layers on top of the wafers to achieve this. The epitaxial process consists of depositing silicon atoms on top of the wafers. These atoms adhere to the surface of the polished wafer in the same crystallographic manner, adding layers of atoms on top of the wafer. Now the entire wafer acts as the seed and we have much better control over the quality of the epitaxial layers. These epitaxial layers are usually very thin, between 1 and 4 μm. The thinner the layer, the less likely it is to have impurities or crystallographic imperfections. In some cases, such as the long wave infrared detectors I discussed in Chapter 4, we need very thick epitaxial layers so the photons crossing the layer have a better chance of being absorbed. We have grown layers as thick as 40 μm, but this isn't easy.

I have concentrated on talking about the silicon process. You can imagine that growing other semiconductor materials, such a GaAs or HgCdTe, is considerably more complex. Nevertheless, the processes are, in theory, about the same.

10.4 Photolithography

Now comes the fun part. How do we fabricate a circuit on these pristine, perfectly flat and clean wafers?

There is an old riddle that asks how you would catch an elephant with binoculars, tweezers, and a jar. Stop reading at this point if you want to think about the answer. You turn the binoculars around and look through them the wrong way. The elephant is now very small, so you pick him up with the tweezers, and drop him into the jar! But seriously, Robert Noyce (1927–1990) (Figure 10.4) at Fairchild had a similar idea, knowing how, from a small photographic negative, using optics, you can create a very large print. His eureka moment was to realize that he could use the same process in reverse, that is, from a large drawing he could make a very small negative and use that negative to fabricate devices.

When I was studying semiconductors in the 1960s, one laboratory exercise was to fabricate a transistor. I used rubylith, a transparent material covered by a red sheet. On a large rubylith sheet ($1 \times 1\,\mathrm{m}^2$) I drew the shape I wanted to project onto the semiconductor material. With an X-acto knife, I removed strips of the red sheet, leaving a pattern with transparent and opaque regions. Using equipment very similar to that found in a darkroom, I created a small ($\sim 1 \times 1\,\mathrm{cm}^2$) negative. I used that negative to fabricate my transistor on a silicon wafer. It was a 5 cm wafer!

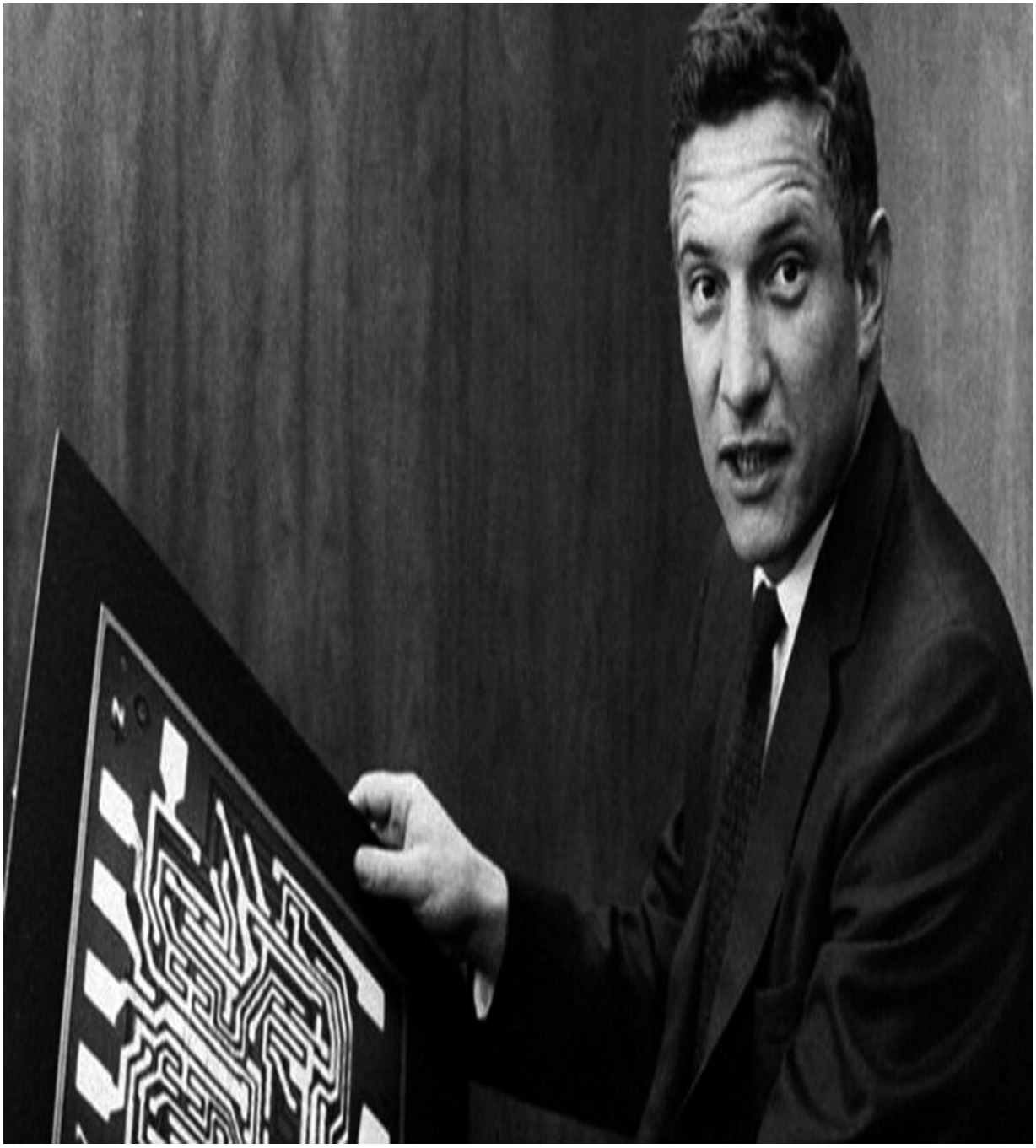


Figure 10.4 Dr. Robert Noyce, observing the photographic process in a darkroom, concluded that the process could be reversed to obtain very small plates from very large drawings.

Source: https://en.wikipedia.org/wiki/Robert_Noyce#/media/File:Robert_Noyce:with_Motherboard_1959.png.

Silicon, in addition to being the second most abundant element on earth, has many other advantages. It is relatively easy to fabricate with the purity needed for semiconducting devices. Gallium arsenide has excellent properties, such as lower noise and higher mobility than silicon, but it is much more difficult and expensive to fabricate very pure wafers from it. Its use is restricted to very specific applications, such as ultra-high radio frequencies, lasers, and ultra-fast electronic switches.

Silicon has also a natural oxide, SiO_2, that is highly insulating (GaAs is not) and has a crystallographic structure similar to that of silicon so that its growth over the silicon does not corrupt the silicon structure. It is also very easy to grow using thermal oxidation ($Si + O_2 = SiO_2$). It grows one layer at a time, very uniformly, so it not only provides insulation, but it also protects the silicon itself. I introduce SiO_2 here because it is an integral part of the fabrication of integrated circuits.

10.5 The Fabrication of a pnp Transistor on a Silicon Wafer

Now that we have almost all the key elements, let me explain how we make integrated circuits. Suppose we want to fabricate a simple pnp transistor. I show the cross-section of the transistor I want to build in Figure 10.5 and the top view of the same transistor in Figure 10.6.

What Figure 10.5 shows is what we call planar silicon technology because the surface of the device is flat. You will recognize the transistor with p (blue) and n (yellow) layers one on top of each other. The lower layer (light blue for p-type) is the collector, which is just the p-type epitaxial layer that we have deposited on top of the supporting wafer. Above the epitaxial layer (yellow for n-type) is the base, and the emitter is the third layer, made of p-type semiconductor (also in light blue), on top of the base. Notice that at the surface we have darker blue and darker yellow islands. To make

good contacts we prefer a highly doped material, p+ (dark blue) and n+ (darker yellow), the + sign indicating that we have highly doped materials. This facilitates the transition of charges between the semiconductor and the metal, and avoids very abrupt junctions. Finally, the black squares on top are aluminum metallic pads that connect this device to other devices or to a pad for bonding wires. These pads allow us to connect other components to the collector (C), the base (B), and the emitter (E). Figure 10.6 shows the aluminum lines that connect this transistor to other devices.

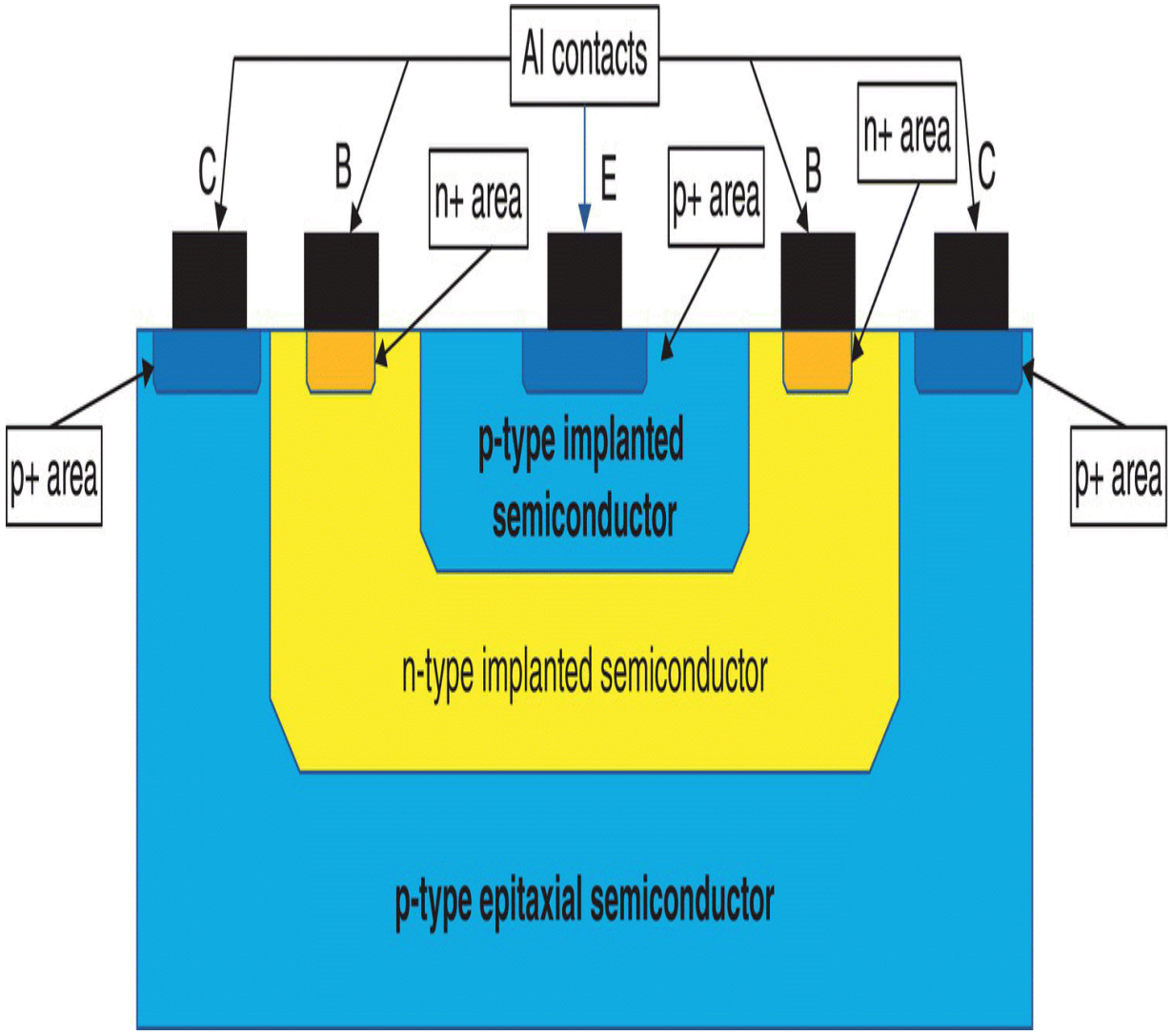


Figure 10.5 Cross-section of the planar transistor we want to build. The p-type collector and emitter (blue) and the base (yellow) are fabricated one over the other, including the contacts (black).

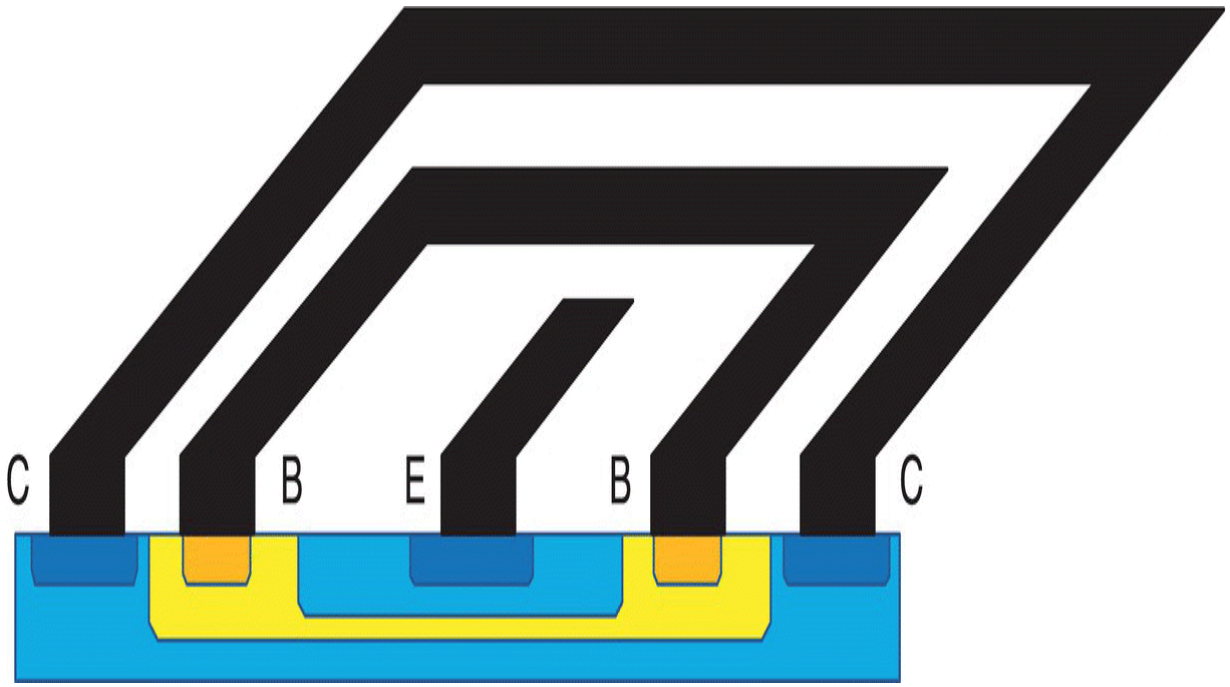


Figure 10.6 Top view of the aluminum lines connecting the different silicon layers.

Now, let's follow the process to fabricate this transistor (Figure 10.7).

We start with a p-type substrate. The thickness of a typical wafer is between 0.3 and 0.8 mm (this is like the dough in a pizza, thick enough to handle all the toppings).

On top of the substrate, we epitaxially grow a much cleaner p-type region with the desired amount of p-type impurities, usually boron. The epitaxial layer is between 1 and 4 μm thick (this is the tomato paste in the pizza analogy).

On top of the epitaxial layer we grow an oxide layer, SiO_2 (in red). To do this, we place the wafer in a furnace and heat it to around 1000 °C while introducing oxygen into the chamber.

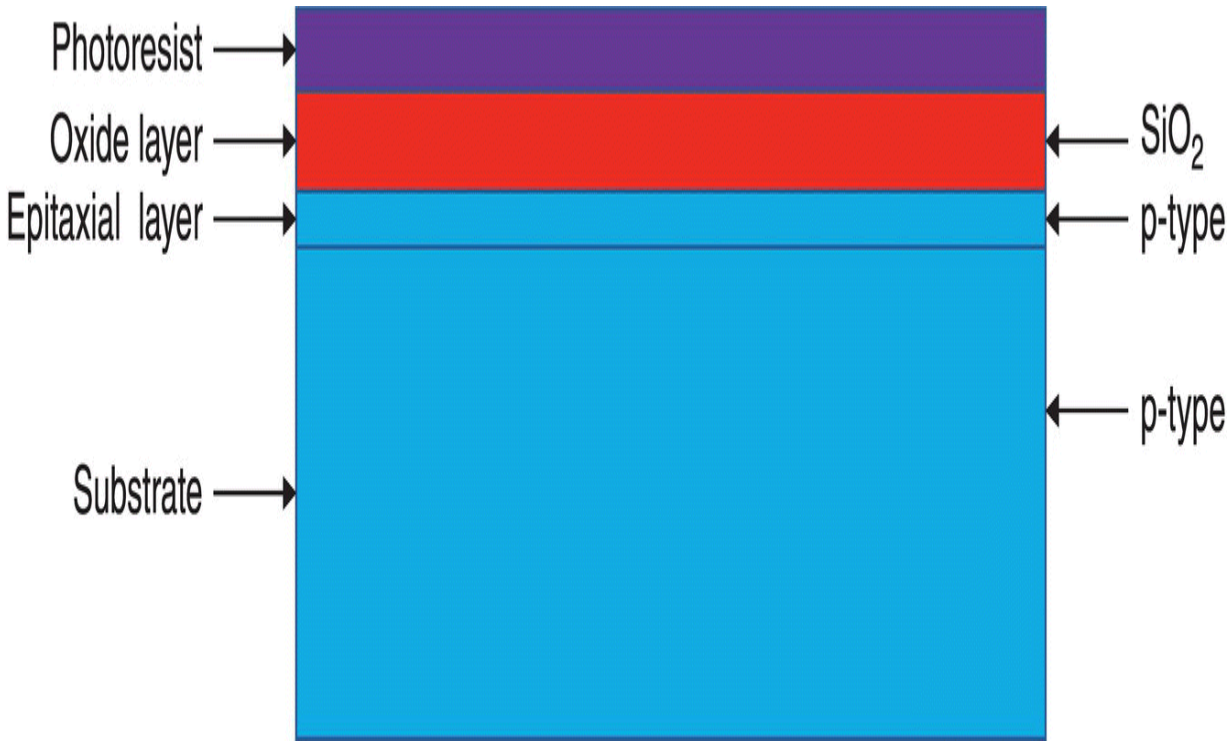


Figure 10.7 First four steps of transistor fabrication: the epitaxial layer on top of the substrate (both p-type, blue) the silicon oxide layer (red), and the photoresists (violet).

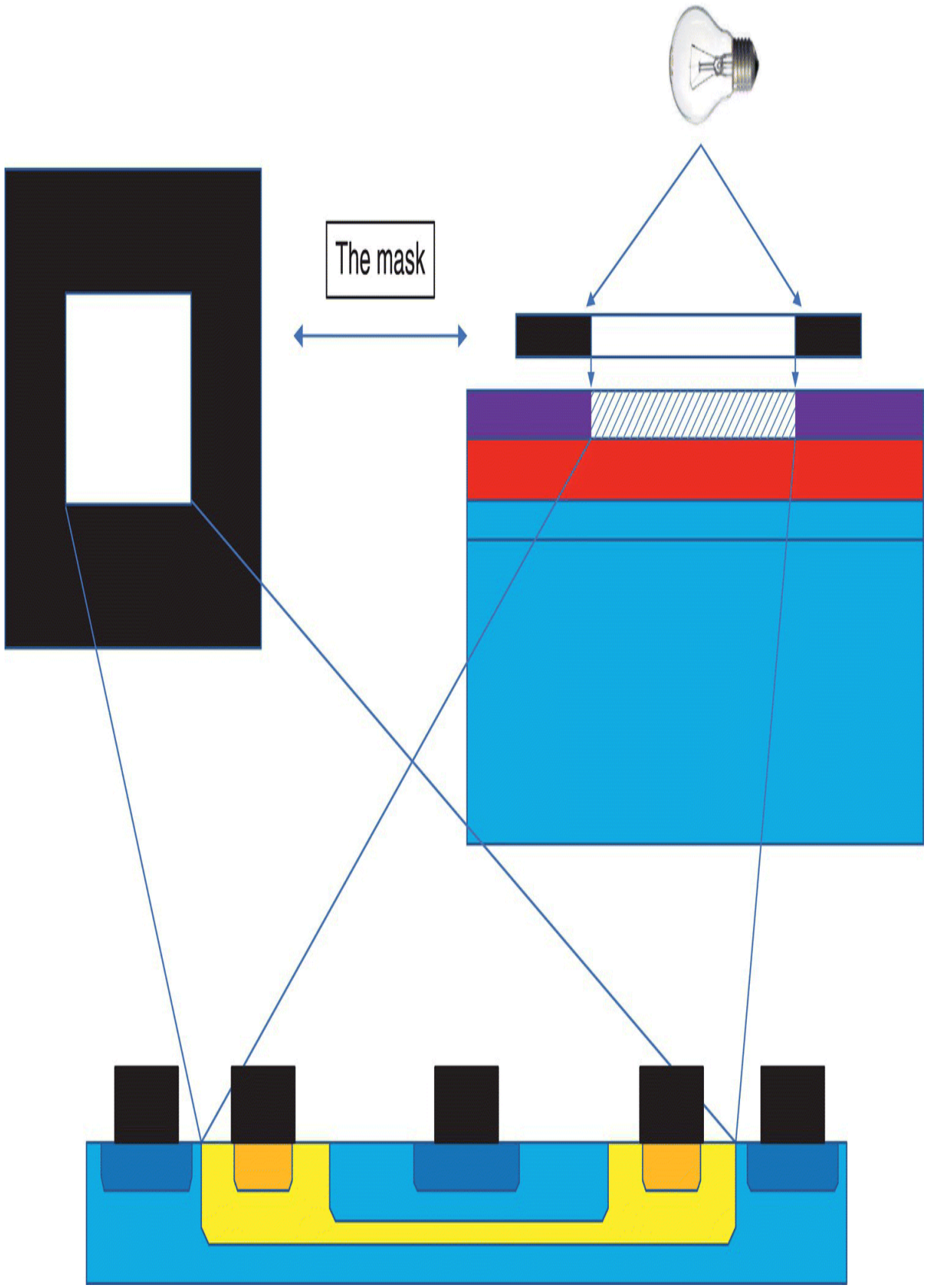

The mask

Figure 10.8 The next step is to photographically illuminate the portion of the wafer we want to uncover. The light acidifies the photoresist, which can then be removed with an alkaline solution.

On top of the SiO_2 we deposit a photoresist (the violet layer in Figure 10.7). The photoresist is a light-sensitive material. There are positive and negative photoresists. The positive photoresist dissolves when we shine light on it. The negative photoresist does the opposite. The negative photoresist has better properties and is used more often. To avoid confusion, I will use only positive photoresist in this chapter.

Now we apply the first masking step (Figure 10.8).

I show the first mask we use at the top left of Figure 10.8. It shows a transparent square surrounded by an opaque area with the exact dimensions as the n-type region (yellow) we want to create. We place this mask on top of the wafer (top right) and shine a light on it. The portion of the positive photoresist that is illuminated is acidified, making it soluble in an alkaline solution. I show the result in Figure 10.9.

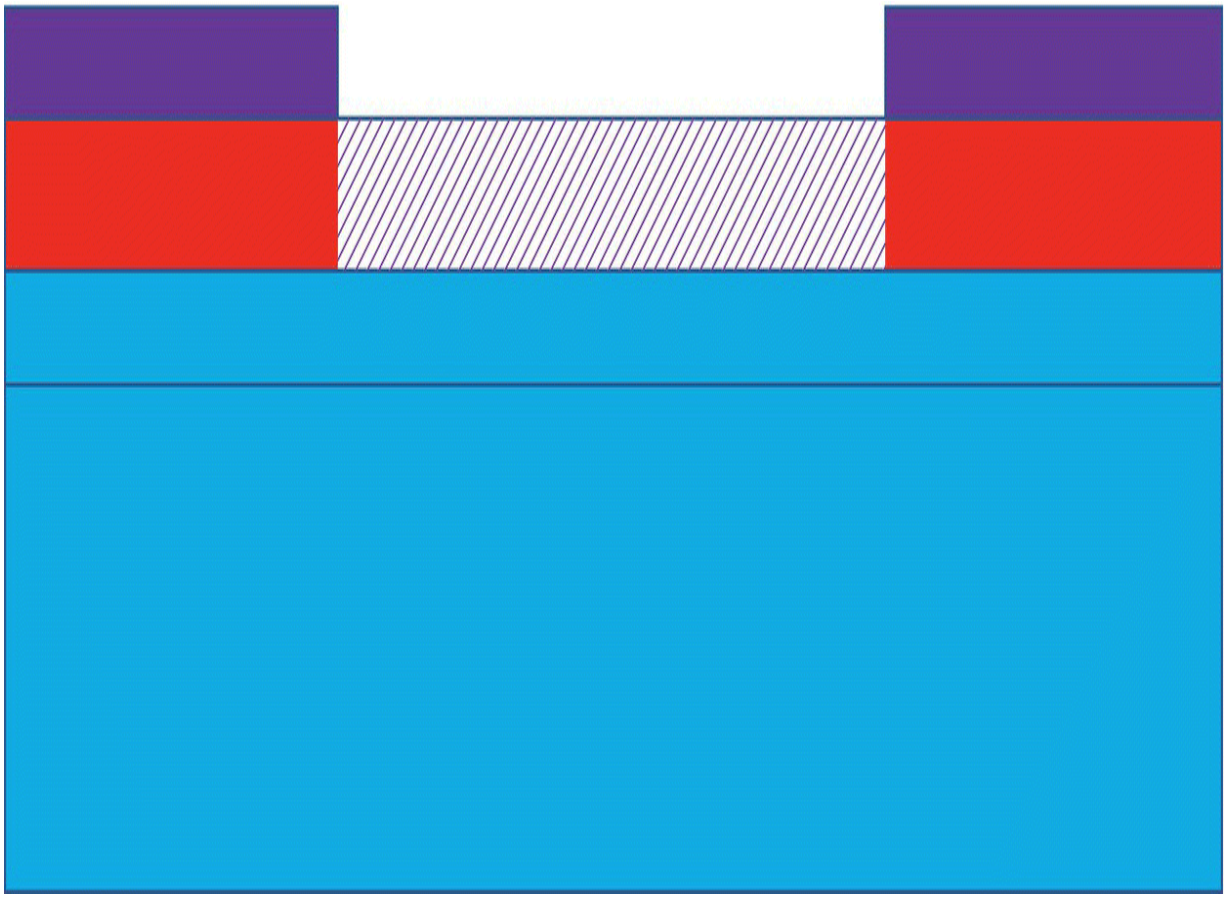


Figure 10.9 The semiconductor after the illuminated part of the photoresist has been removed.

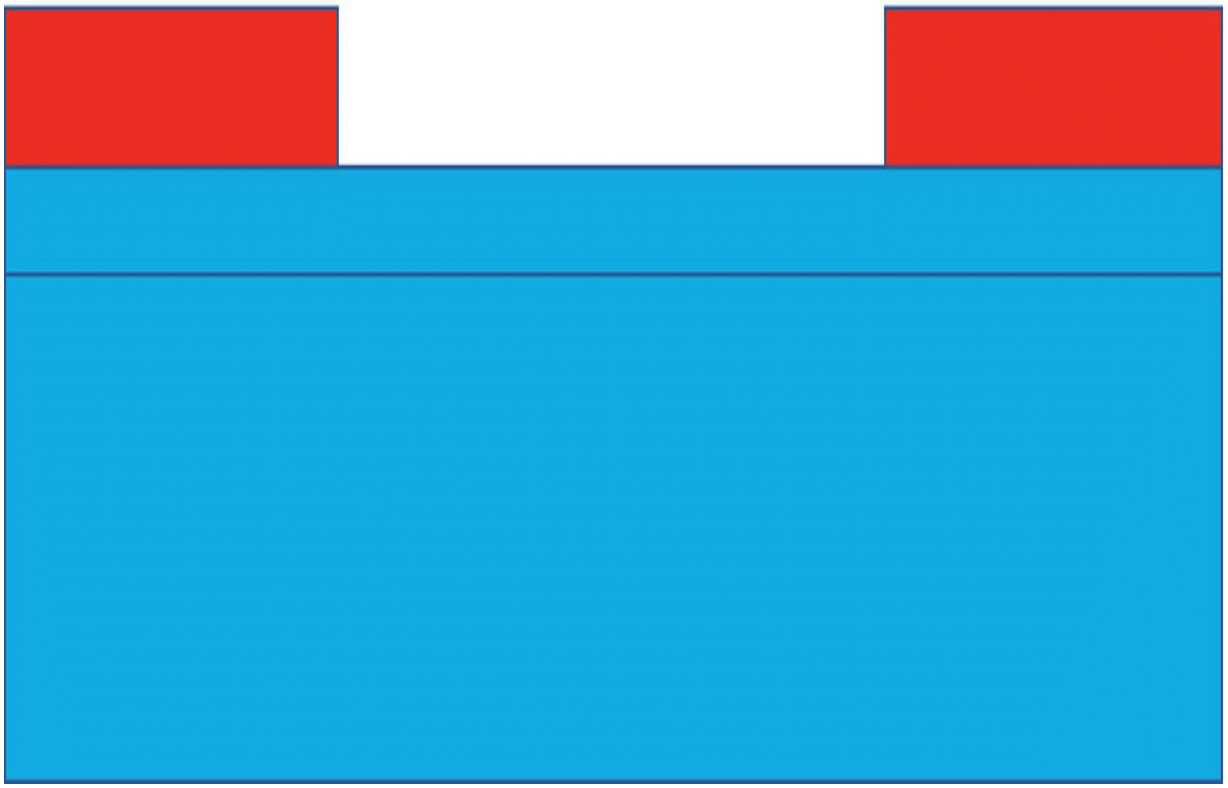


Figure 10.10 We remove the oxide with ammonium fluoride and the excess photoresist on top of the oxide, leaving a clean opening in the epitaxial semiconductor surface.

Now we remove the oxide under the open photoresist window with an etching solution such as ammonium fluoride. After we clean and remove the extra photoresist (Figure 10.10) we are ready to fabricate our n-type base.

10.6 A Digression on Doping

Now that we have the photoresist window open, I want to discuss the ways we can locally change the electrical characteristic of the semiconductor by doping, that is, adding the desired impurities. This is different from the boule, where we wanted the entire boule to have either p- or n-type impurities with a very uniform concentration. Now we want to just affect the region that we have opened. We want to change the impurity concentration under the

window from p to n. There two ways of doing localized doping: thermal diffusion and implantation.

10.6.1 Thermal Diffusion

In the thermal diffusion case, we place the wafer in a chamber, heat the wafer, and introduce the gas we want to use to dope the open region of the wafer. Figure 10.11 shows the case where we want to add n-type impurities into the open area using diffusion.

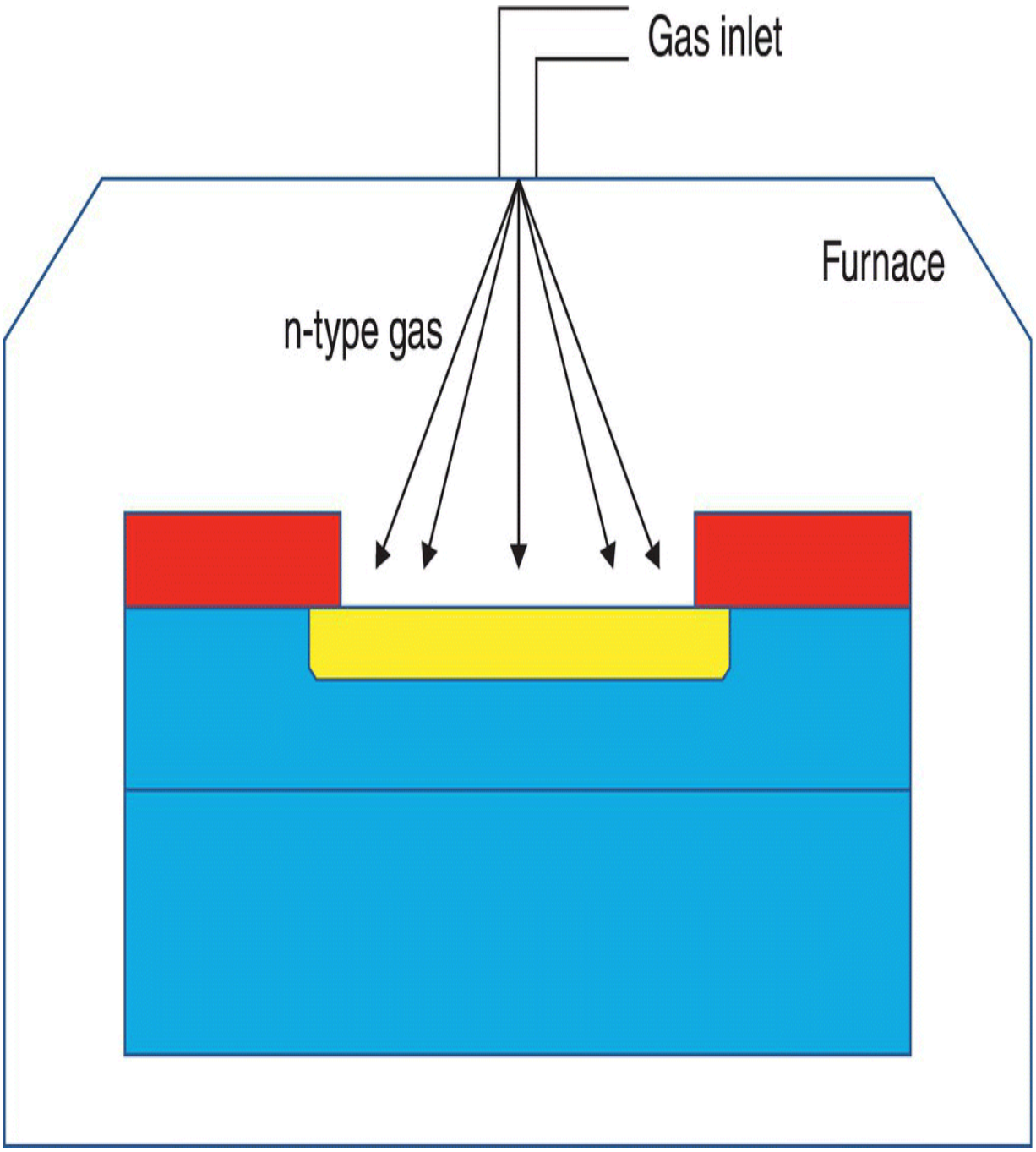


Figure 10.11 The semiconductor, with the desired oxide removed, is located in a chamber, the specific gas is injected at high temperature, and the impurity atoms diffuse into the p-type epitaxial layer creating an n-type region (yellow).

When we use the diffusion system, the number of n-type impurity atoms is not constant in each atomic layer. We are always going to

have more impurity atoms at the top surface than inside the epitaxial layer, but, in general, that is not a problem since we want an np-type junction between the p-type epitaxial layer (blue) and the n-type base (yellow). Also, because of the way diffusion works, impurities will diffuse under the oxide (red). How far the n-type impurities go inside the p-type epitaxial layer depends primarily on the temperature of the wafer and how long the wafer stays under the gas. The higher the temperature the faster the impurity atoms diffuse into the wafer, and the longer it is exposed the farther the impurity atoms will go in. Typical n-type impurities are phosphorous, arsenic, and antimony, which are all valance five materials, but for p-type material the doping is restricted to boron.

This is not the only time that we will place the wafers in a hot oven and as we go through different thermal cycles the diffused impurities will tend to move and diffuse still more. Figure 10.12 shows this effect.

Curve a in Figure 10.12 shows the distribution of impurities after we finish the deposition, the largest concentration at the surface, but after many other subsequent heat treatments the impurities at the surface decrease (curve b) and the impurity atoms diffuse further into the p-type epitaxial layer. The total number of impurity atoms is the same, but the distribution is different. The process engineer must consider these future heating steps when calculating how much time and what impurity flow to use.

10.6.2 Implantation

The second doping method is implantation. As components in an integrated circuit become smaller and smaller, we need to find ways to insert controlled impurities in tinier and tinier spaces. As I mentioned before, the diffusion method sends impurities under the oxide. Thus, we need sufficient space between devices to ensure that none of the impurities in one device migrate under the oxide and make contact with an adjacent device. Additionally, the majority of the dopants are at the surface (see Figure 10.12), where there

are the most imperfections and unwanted impurities. Figure 10.13 shows a very simplified diagram of an ion implanter. The idea of the ion implanter is the same as someone shooting a BB gun into a soft target. The speed, or kinetic energy, of the BBs determines the extent to which they penetrate the target.

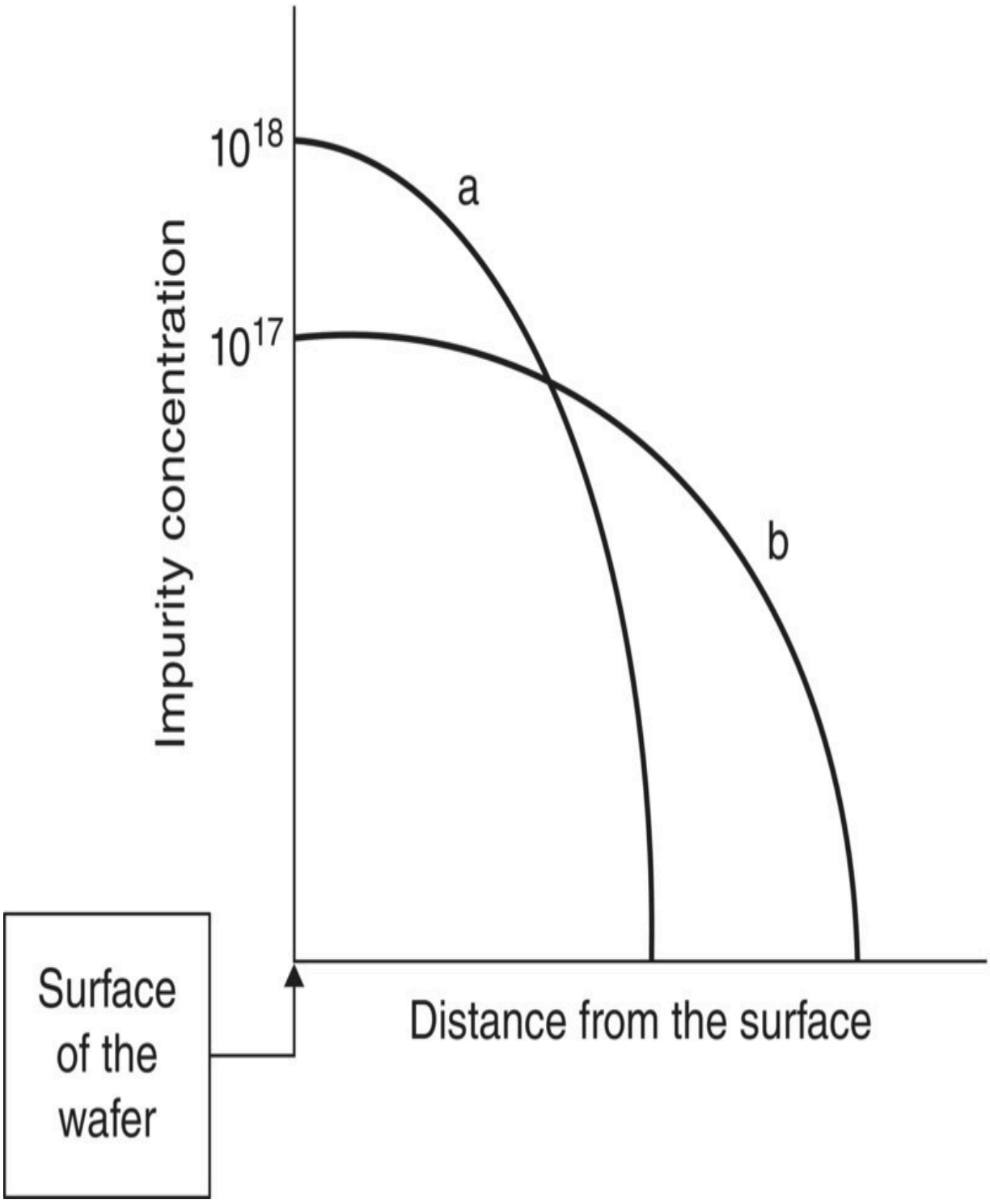

10^{18}
10^{17}
a
b
Impurity concentration
Distance from the surface
Surface of the wafer

Figure 10.12 The impurity concertation at the end of the deposition (curve a) is largest at the surface and decreases as we move away from the surface. After many additional thermal steps, the impurity diffuses more into the epitaxial layer (curve b), although the total number of impurities remains the same.

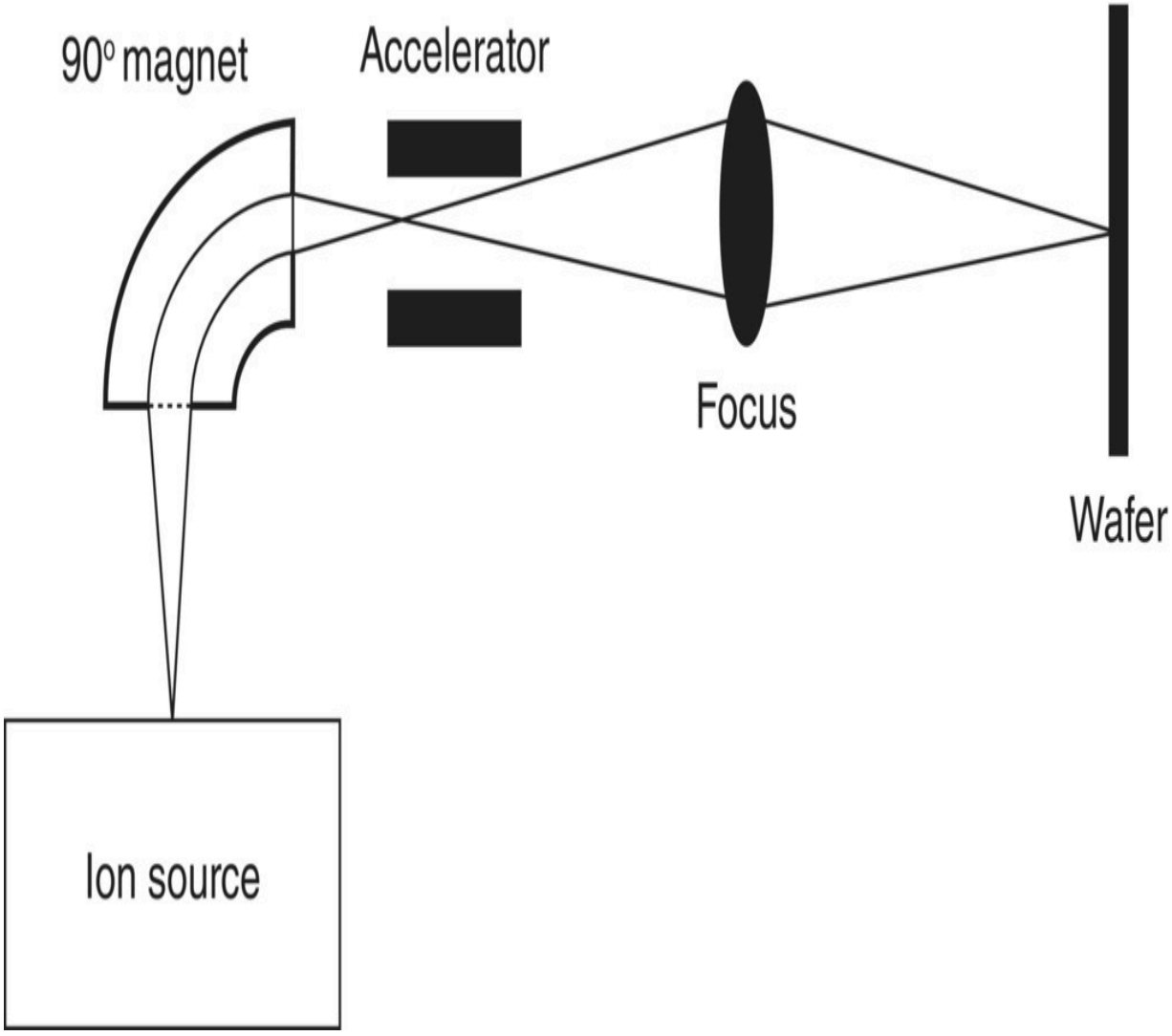


Figure 10.13 An ion implanter consists of an ion source, a magnet to separate and block unwanted impurity atoms, an accelerator, and a focusing system so the beam can hit the wafer at the desired place and at the desired speed.

The implanter consists of a source of ions. Ions are charged (positive or negative) atoms. We want only the p- or n-type impurities to be

deposited into the wafer. The ions are sent through a magnet that, because atoms have different masses, separates the different ions, letting only the atoms we want go through and blocking the unwanted ones. This 90° magnet is the same as a mass analyzer. Depending on the charge and mass of the particles, they bend at different angles. The unwanted ones hit the wall of the mass analyzer.

The next step is to accelerate the ions and focus them onto the region we want to implant. We accelerate the ions from thousands to millions of electron-volts depending on the type of implanter. We scan the beam so that the entire wafer is covered and all the oxide-free surfaces in the wafer are implanted. The wafers sit in holders that move, so the scanning can be done mechanically by moving the wafers or optically by changing the direction of the beam, or both. When one wafer is completed, another automatically moves into place.

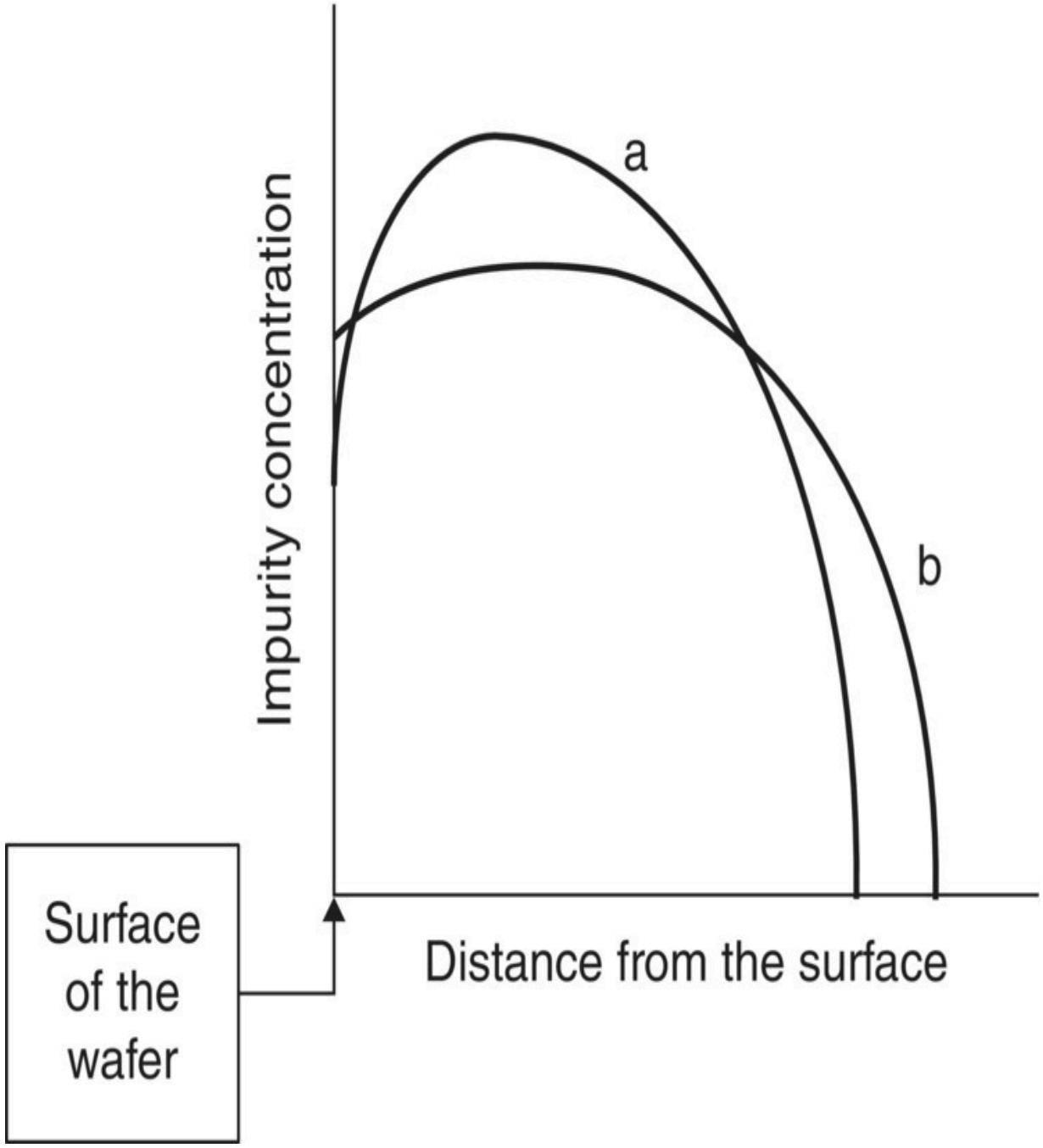


Figure 10.14 The impurity concentration in an implanted wafer as a function of distance shows that the maximum concentration is inside the semiconductor after implantation depending on the beam energy (curve a). After later heat treatments the impurity concentration diffuses, as I show in curve b.

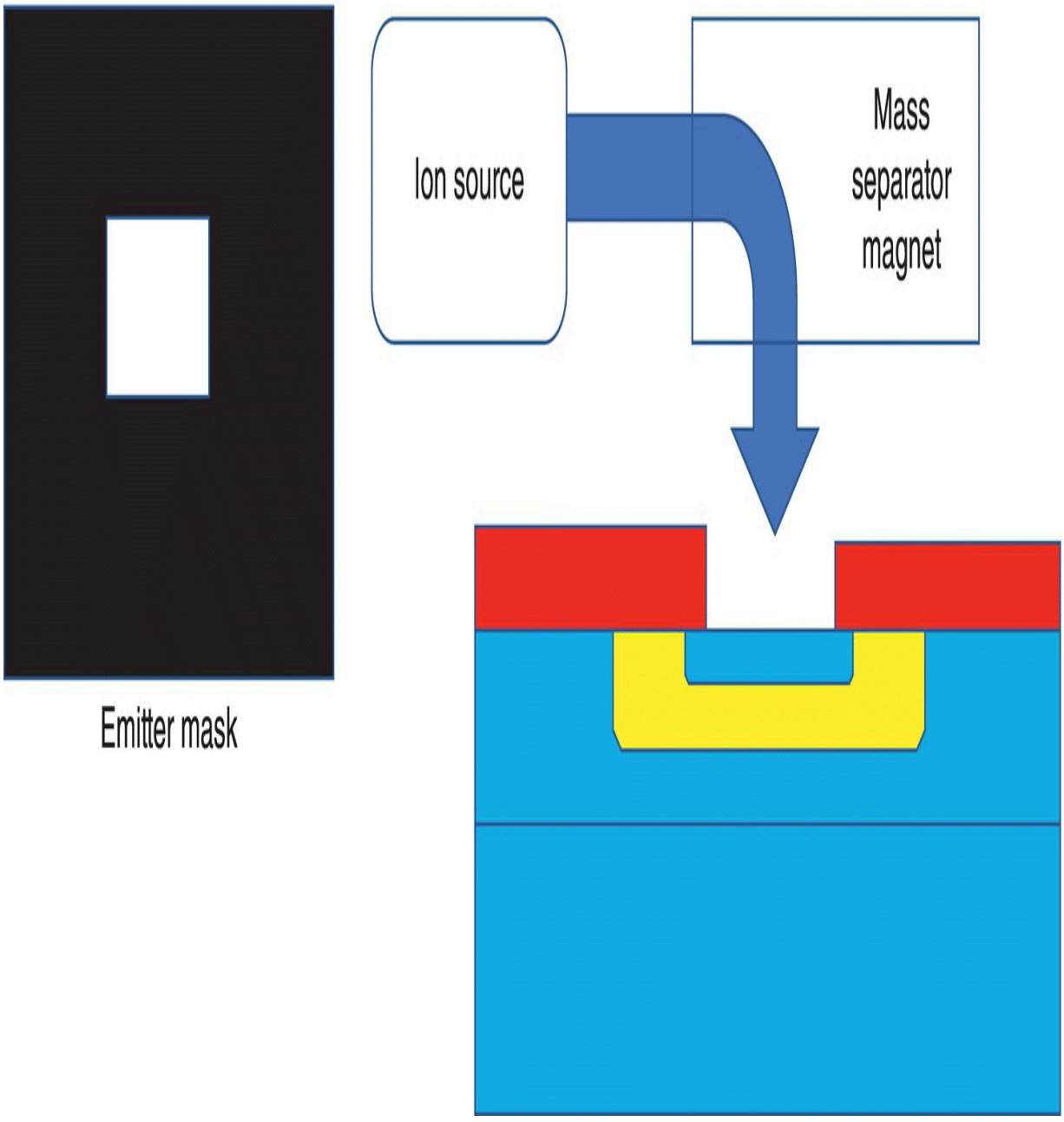


Figure 10.15 Using an ion implanter we fabricate the emitter region using a mask with a smaller opening than the one we used for the base.

The distribution of impurities in an implanted wafer is different from that in the diffuse method (Figure 10.14). As you would expect, the highest concentration of dopant impurity is no longer at the surface but inside the material, depending on the beam energy (curve a). After any of the subsequent heat treatments, the concentration

diffuses into the body of the transistor and the distribution flattens out (curve b). The total number of impurity atoms remains the same.

The bombarding of the surface of the semiconductor with high-speed atoms damages the surface. The final step is therefore to anneal the surface, which we typically do while heating it at between 700 and 1000 °C. This is considerably lower than the melting point of silicon (1414 °C) so the surface quality is restored without diffusing the impurities under the oxide, as happens with the diffusion method. This is one more reason why implantation is the preferred method.

Using an ion implanter, we fabricate the p-type emitter on top of the n-type base. The photolithographic process is the same as explained before (Figure 10.15).

The opening in the mask (on the left) is now smaller than the one we had for the base, defining the smaller p-type region that we want inside the base. As we did with the n-type diffusion of the base, we first grow the oxide on the entire wafer and then cover it with photoresist. We place the mask on top of the wafer, illuminate it, remove the acidified photoresist, and remove the exposed SiO_2 and the remaining photoresist, with developer and etching. This opens the region we want to change to p-type, this time using an ion implanter. This completes the fabrication of the transistor.

10.7 Resume the Transistor Processing

10.7.1 The Contacts

Now we need to make connections to the outside world. We know that the contacts work better if they are connected to very heavily doped regions so that there are no abrupt transitions from a lightly doped material to a highly conductive metal. We therefore want to open spaces up to implant very heavily doped impurities. Figure

10.16 shows the two masks we use to implant the highly doped contact regions.

I will not repeat the explanation over and over, but you can see what the process is: grow oxide, apply photoresist, illuminate and develop the photoresist, remove the acidified photoresist and the oxide to leave the holes, and implant the exposed areas. I show the heavily doped implanted contacts (darker blue and darker yellow) in Figure 10.17. As I mentioned at the beginning of this chapter, the process is repeated as many times as we need.

10.7.2 Metallization

The final step in our process is metallization. Leaving the contact holes open we now cover the entire wafer with a thin layer of metal, the black layer in Figure 10.17. Figure 10.18 shows the mask layer we use to define the conductive lines.

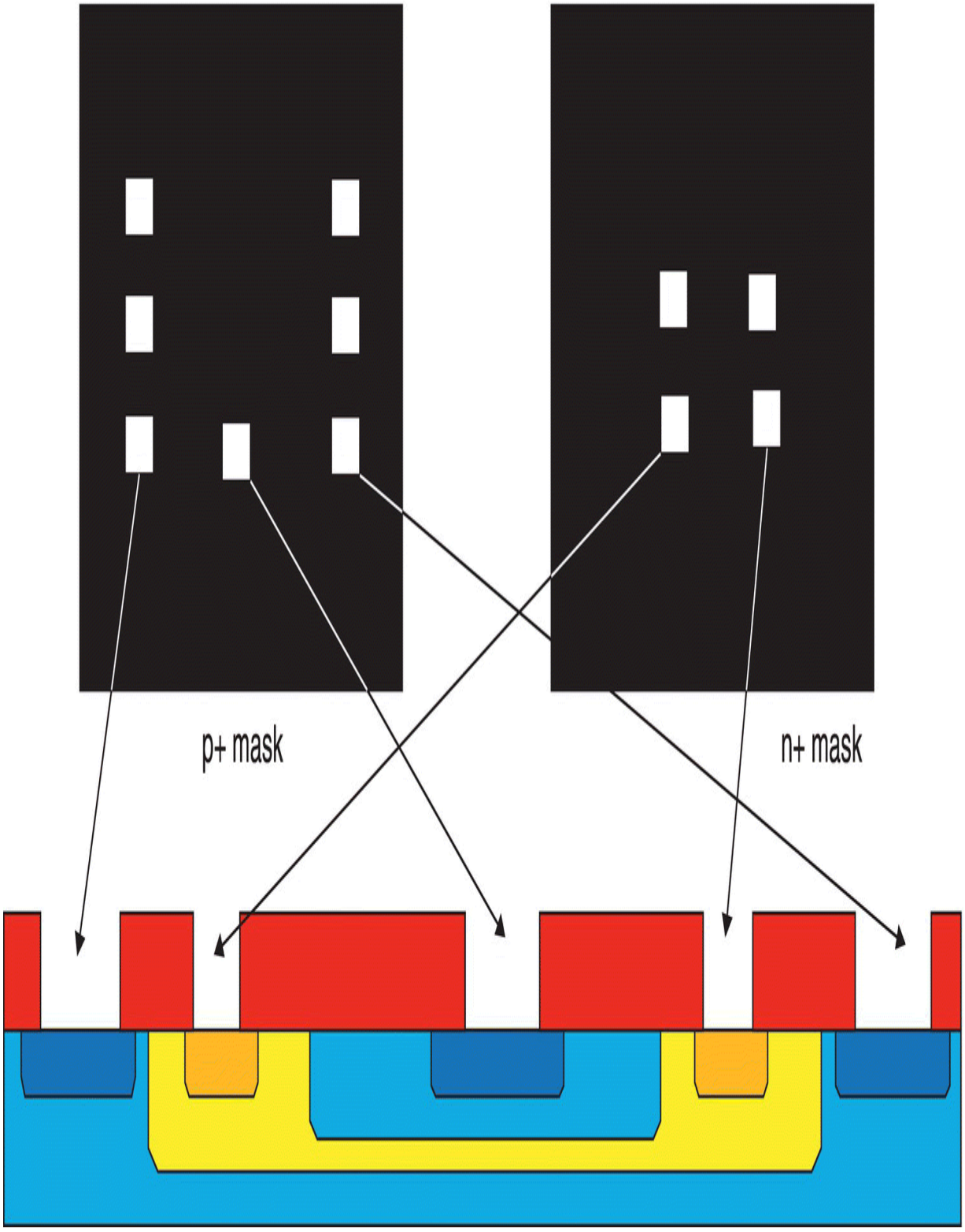


Figure 10.16 Mask used to create the p+ (left) and n+ (right) regions.

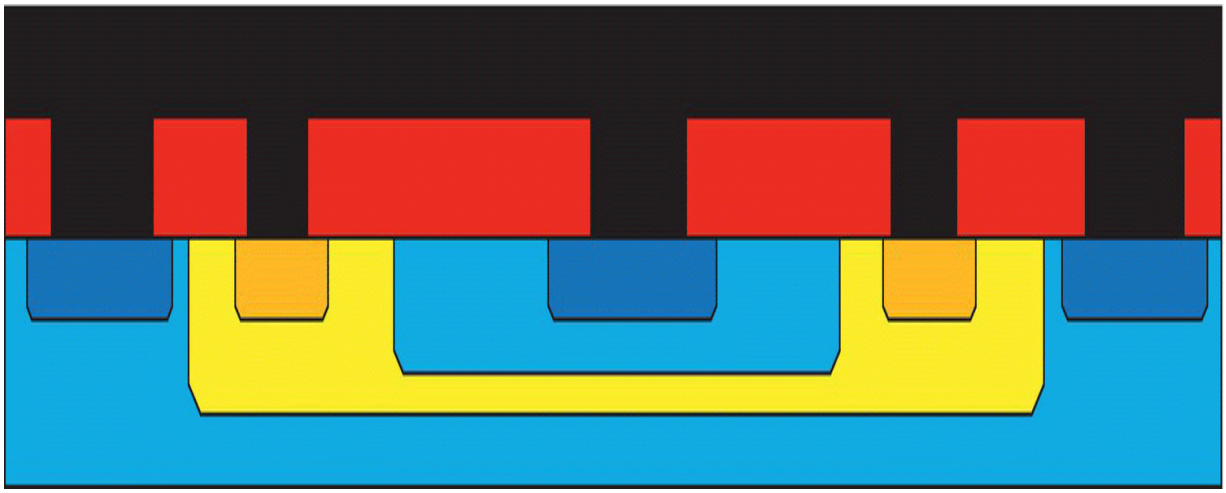


Figure 10.17 A wafer covered with a metal layer makes contact with all the n+ and p+ regions.

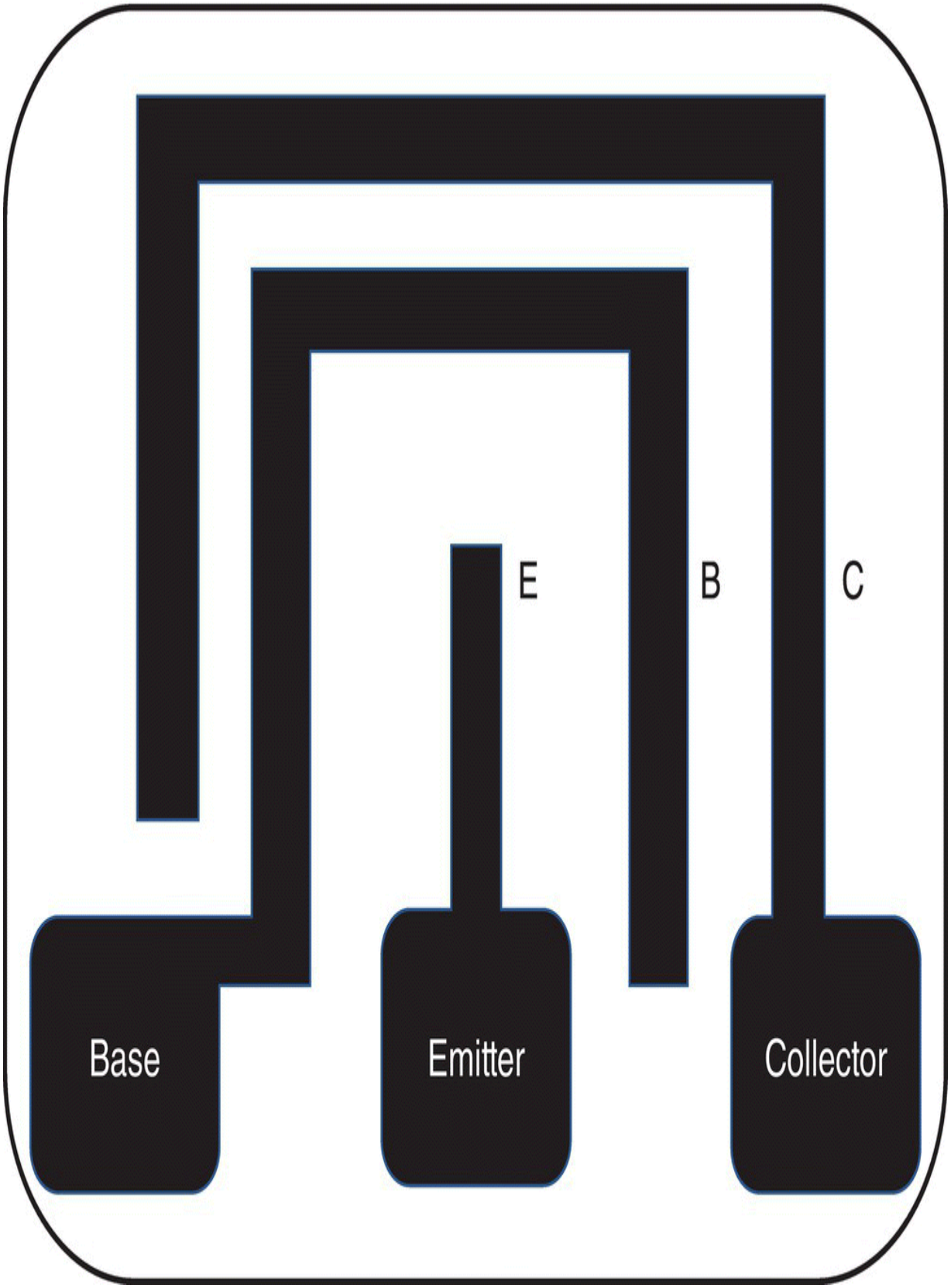

E
B
C
Base
Emitter
Collector

Figure 10.18 The aluminum mask connects each contact on the wafer to areas where we can solder external contacts.

The aluminum mask not only connects all the contact points going to the same transistor function, but also provides a larger region so we have enough area to solder the external wires.

10.7.3 Multiple Interconnects

One final thing on interconnects. Right now, we can fabricate as many as one billion transistors in a 1-cm^2 chip, about a quarter of a size of a standard postage stamp. As Larry Zaho mentions in an article on interconnects (https://semiengineering.com/all-about-interconnects), to make contact with all of these devices requires an equivalent of 30 miles of wires. We can't possibly do this with a single aluminum layer. Figure 10.19 shows some of these levels of interconnection.

At the bottom there are the individual transistors and other electronic components. Above the transistors are the local interconnects. These are usually thin and short lines. Above those, there are the global interconnects, thicker and longer. Each layer requires the repetition of the same metallization process I explain above. Not all the layers are metallic. Some of them use highly doped polysilicon.

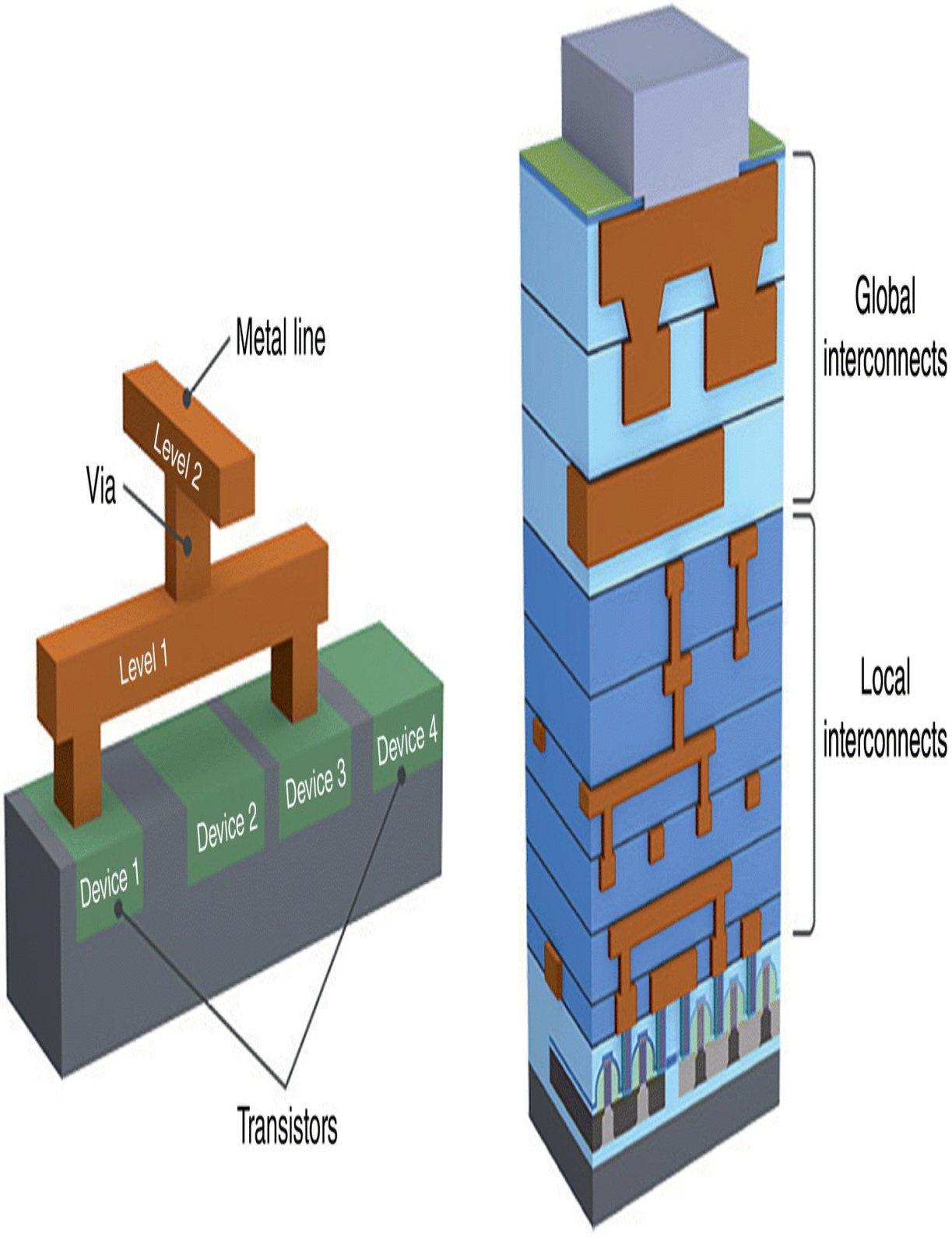

Metal line
Level 2
Via
Level 1
Device 1
Device 2
Device 3
Device 4
Transistors
Global interconnects
Local interconnects

Figure 10.19 Modern electronic integrated circuits use multiple levels of interconnection.

10.8 Fabrication of Other Components

In order to bias and use the transistors, we need also resistors and capacitors. We use the same techniques to fabricate them as we use to build transistors.

10.8.1 The Integrated Resistor

Figure 10.20 shows the process of fabricating a resistor.

Figure 10.20A shows the cross-section of a resistor. It looks very much like the transistor that we have already fabricated. We use the n-type region (yellow) as an isolation island. On top we implant, as we did with the transistor, a p-type region (light blue) and two p+ regions at the two ends of the top p-type region, what used to be the emitter contact (dark blue). We cap this with aluminum contacts. We have fabricated a resistive path isolated from the epitaxial substrate (Figure 10.20B). The value of this resistor depends on how high, or low, we dope the p-type region. If we want a high resistance, we use light doping. We do the opposite, high doping, if we want low resistance, and use low doping. We also have the ability to change both the thickness and the length of the line, as I show in Figure 10.20C, making the resistance three times higher in this particular example.

It is also common to fabricate resistors using a polysilicon layer, but this is a different process. We have to deposit a polysilicon layer and then etch as we did with the metal. We already use polysilicon layers in some multilevel connecting layers, thus we do not need an extra step to use one of these layers as a resistor.

There is also another way of fabricating a resistor by using an epitaxial layer which sits on top of the substrate. Even though both are p-type materials, the substrate almost always has a much higher resistivity than the epitaxial layer. The trick then is to etch in the

very thin epitaxial layer a trench all around the resistor to define its area (Figure 10.21).

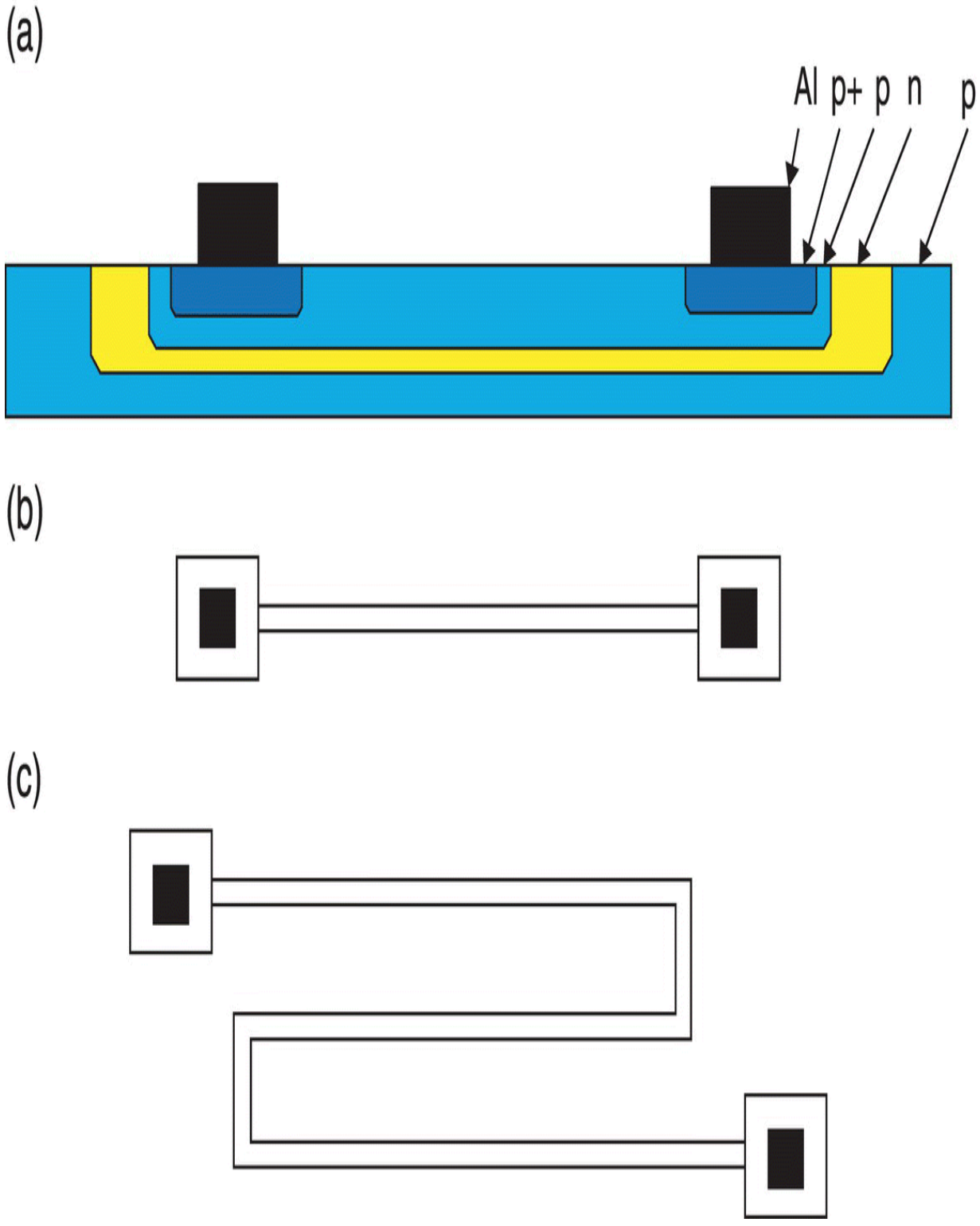


Figure 10.20 Fabrication of a resistor. Cross-section of an integrated resistor (A) and different shapes and forms the resistor may take to increase or decrease its resistance (B and C).

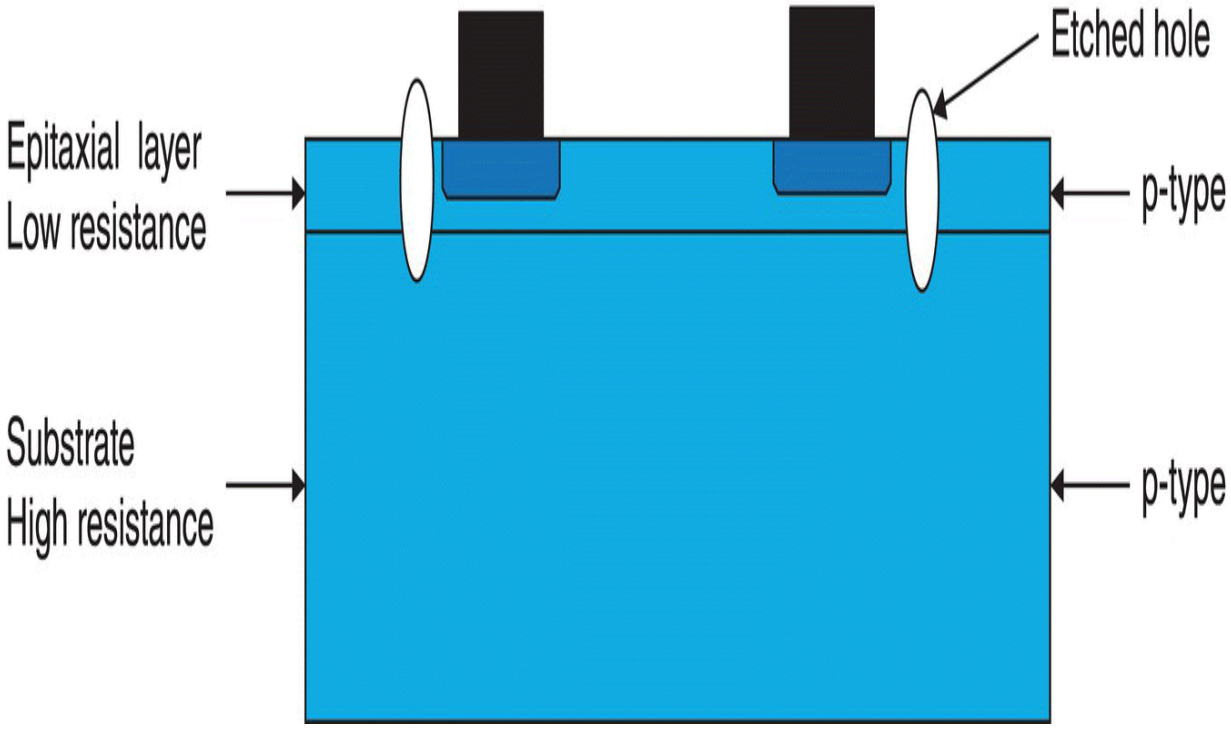


Figure 10.21 Another way to fabricate a resistor is to use the epitaxial layer with etched spaces to delineate the resistor.

At the two ends of the isolated area, as we have done before, we create a p+ contact and a metal contact. Isolation etched holes (and now I am talking about real holes, not electronically charged holes) are also used quite often to isolate other devices like transistors.

10.8.2 The Integrated Capacitor

It is also very easy to fabricate capacitors (Figure 10.22). An integrated capacitor consists of a piece of metal (black) on top of a heavily doped semiconductor (dark blue), separated by an insulating SiO_2 layer (red), the same layer we have been using to define the etching regions. Notice now that the entire area under the oxide is highly doped since we want this area to act as the capacitor's second metal plate.

10.8.3 The Integrated Inductor

There is no easy way to fabricate inductors using the integrated process. The higher the frequency, the smaller the inductor needs to be. For very high frequencies, well above 1 GHz, we can use spiral patterns, which can be imbedded into a chip (Figure 10.23). The most common shape is the rectangular (Figure 10.23A) or circular spiral. Notice that we need two levels of metal since the contact to the right of Figure 10.23A has to cross, but not short, the metal lines that form the spiral. Figure 10.23B shows circular metal lines with a crossover on top instead of the spiral formation, and this can have several metal circles, not just the two I show. This provides a more symmetrical magnetic field. Finally, we can also have several layers, as I show in Figure 10.23C, so we can have a higher inductance in a smaller space. Since the present processing lines already have four or five or more levels of metallization, this process of multiple layers increases the inductance without occupying too much space. Most mobile phones work with frequencies between 0.8 and 2.1 GHz, so they may use one of these imbedded inductors. For lower frequencies the inductors are too big to be integrated with other electronic devices.

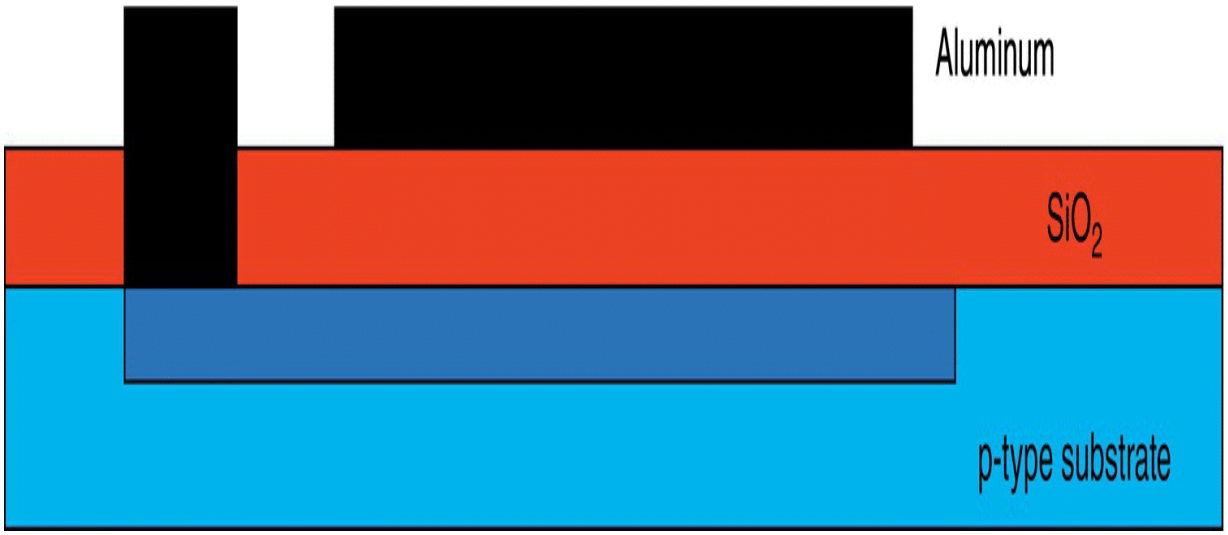


Figure 10.22 Capacitors are fabricated using the same techniques as MOSFETs with a heavy doped semiconductor and a metal plate separated by the SiO_2 insulator.

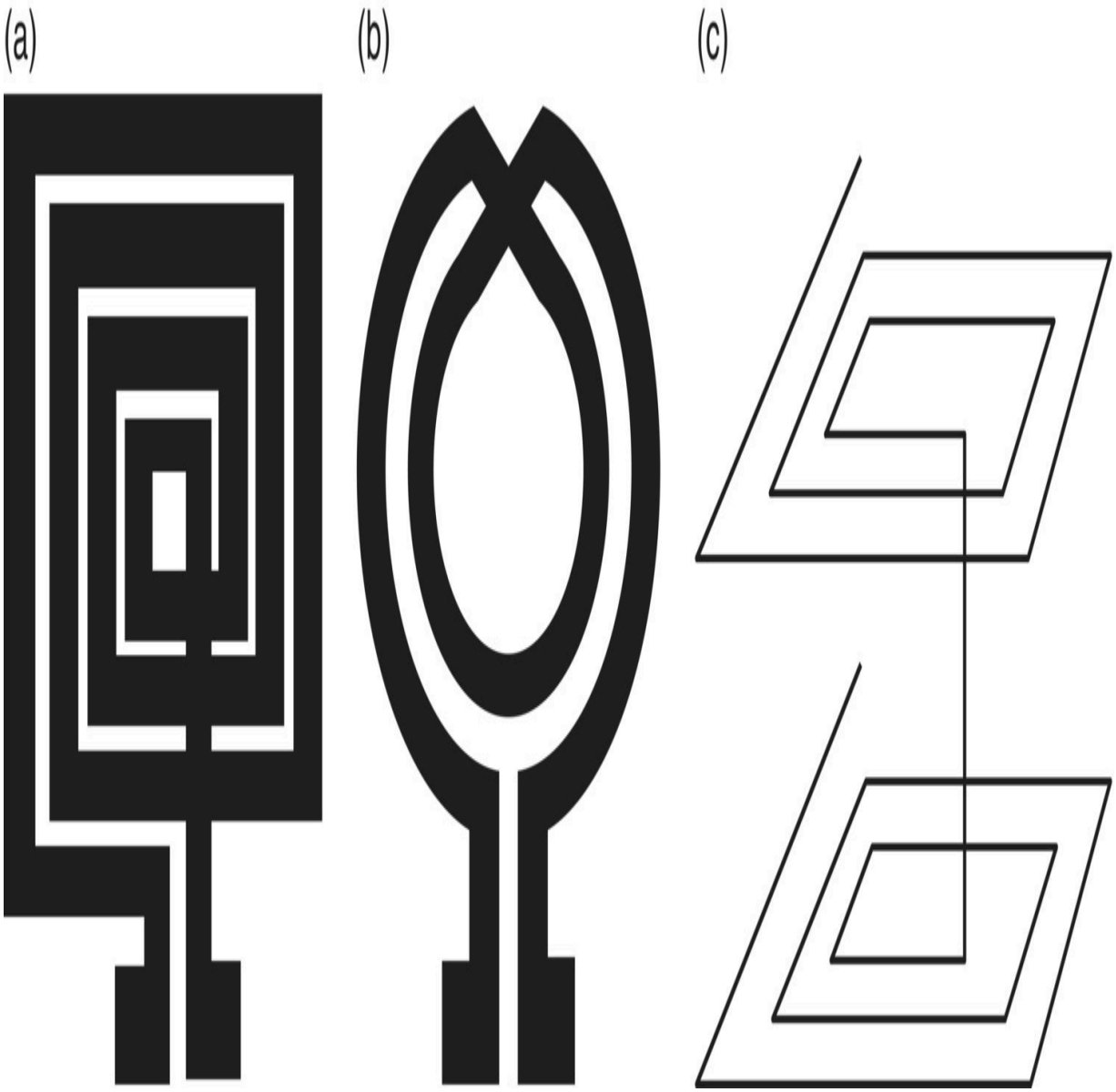


Figure 10.23 Spiral inductors can also be fabricated in a spiral form, as a rectangular spiral (A), as a circular crossover (B) or in multiple levels (C).

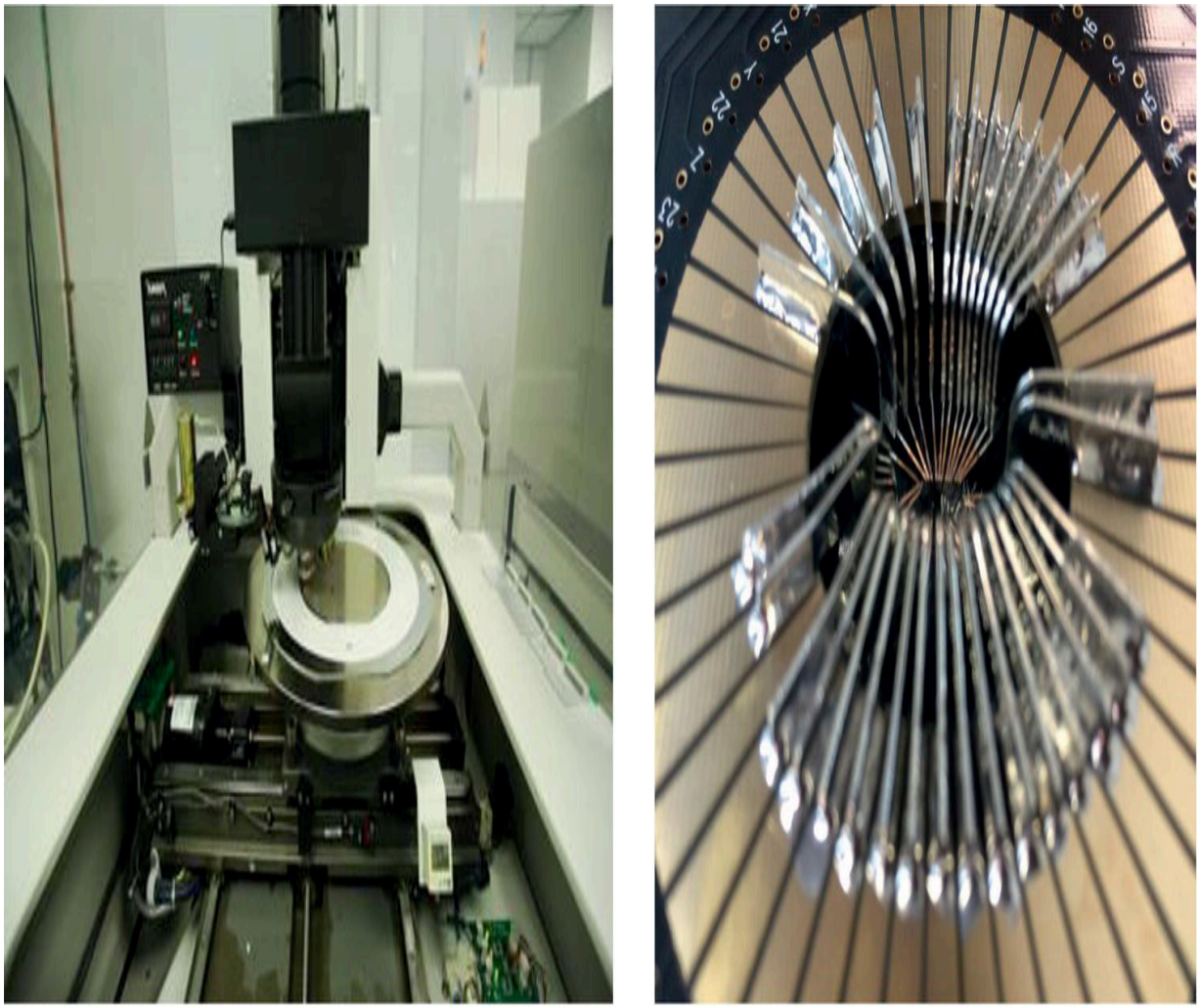
10.9 Testing and Packaging

Once the wafers have completed the processing, we have to test each die to see that there are no defects or imperfections that makes the die inoperable. I show a full processed wafer in Figure 10.24 and in Figure 10.25 I show a probe tester. The tester holds

the wafer and some very tiny probes (right of Figure 10.25) that connect the tester to the important pads of each die that determine its functionality. If the die is defective, the tester deposits an ink dot on top of the specific bad die.

Figure 10.24 A fully processed wafer.

Source: https://en.wikipedia.org/wiki/Wafer_(electronics)#/media/File:ICC_2008_Poland_Silicon_Wafer_1_edit.png.



Micross components

Figure 10.25 A modern probe tester (left) and the very thin conductive probes that connect to the die pads and test the performance of the array (right).

Source: https://www.micross.com/electrical-test/wafer-test/ (left); http://www.alphaprobes.com/gallery.html (right).

The next step is to saw the wafers with a diamond saw and separate the dice. We discard the dice that have a dot and place the good

dice on a trade. They are them inspected and sent for packaging.

There are many different ways of packaging these electronic devices. We have to consider many properties that determine which is the best packaging, the number of leads, the heat generated, the vulnerability to heat or radiation, and the physical protection of these delicate chips.

Figure 10.26 shows the packaging of single devices like transistors and diodes. The typical single transistor package (Figure 10.26A and B) has the three leads for the emitter, base, and collector. Figure 10.26C and D shows packages for power devices. Both can be bolted into a metallic chassis that helps dissipate the heat. Notice that the power package in Figure 10.26D has only two leads for the emitter and base contacts. The body itself is the contact to the collector, which is grounded.

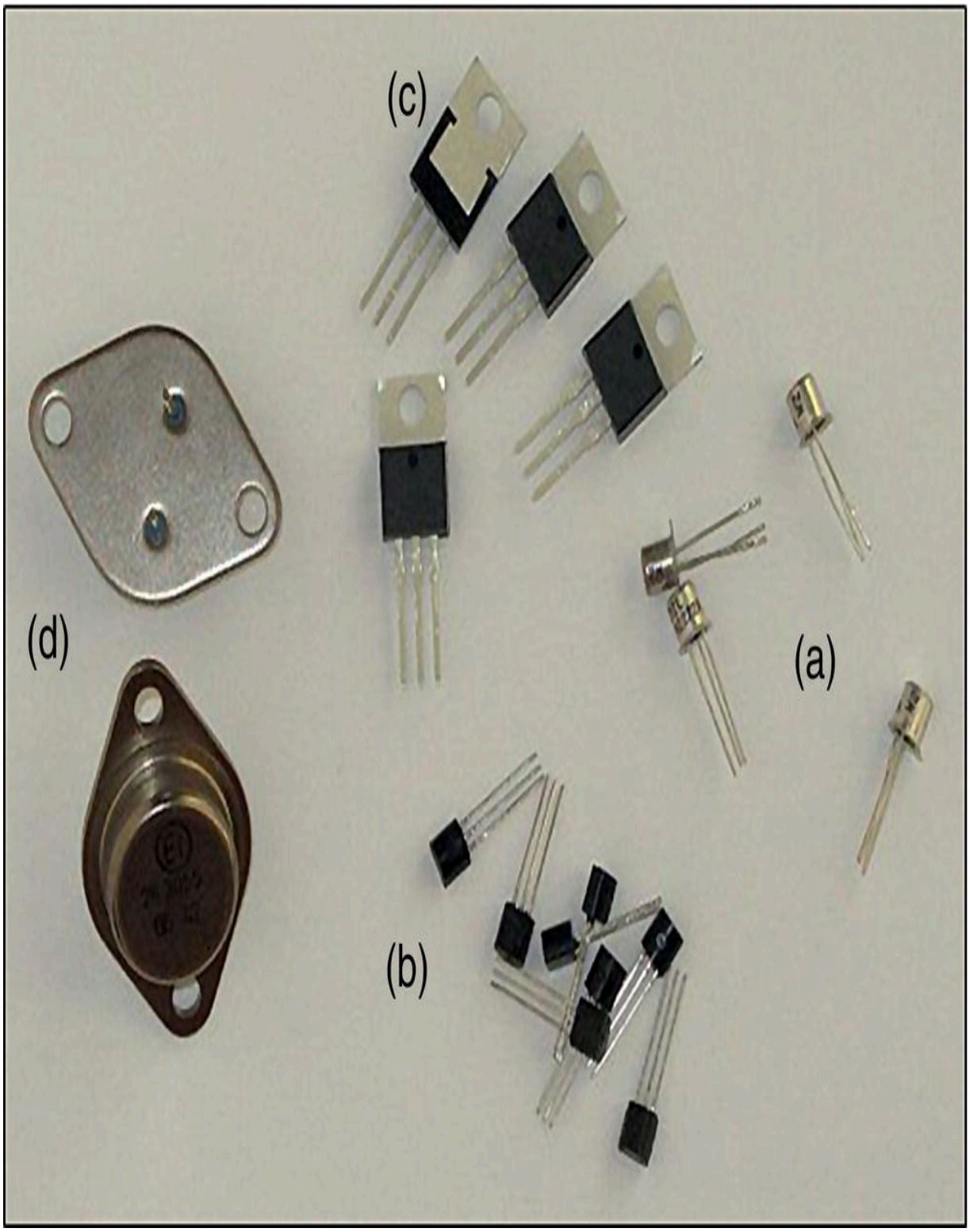

(c)
(d)
(a)
(b)

Figure 10.26 Single electronic device packaging with the three inputs for emitter, base, and collector (A and B). The two packages on the right C and D are power packages that can be bolted to the chassis for good heat removal.

Source: https://www.pinterest.com/pin/440649144770494687.

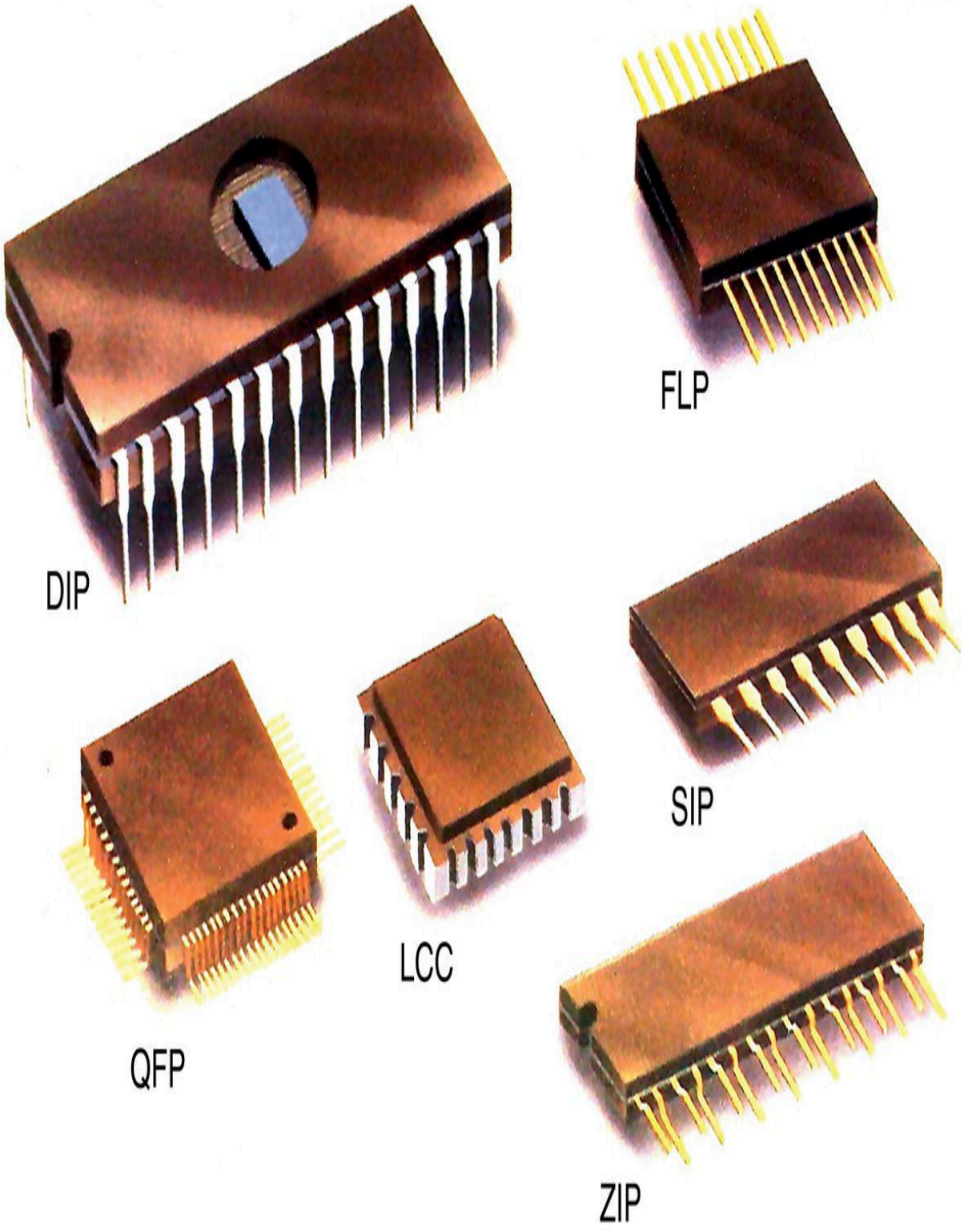

DIP
FLP
SIP
LCC
QFP
ZIP

Figure 10.27 In a flat package the chip sits in the middle and is bonded to the legs of the package.

Source: http://www.chipsetc.com/integrated-circuit-package-types.html.

As the complexity and number of inputs and outputs increases, packaging has also become more complex. Figure 10.27 shows a very common way of packaging electronic components such as operational amplifiers. The chip is firmly attached at the center of the package and then thin aluminum or gold wires connect the chip's pads to the external leads. We can then place the package on a motherboard and solder the leads.

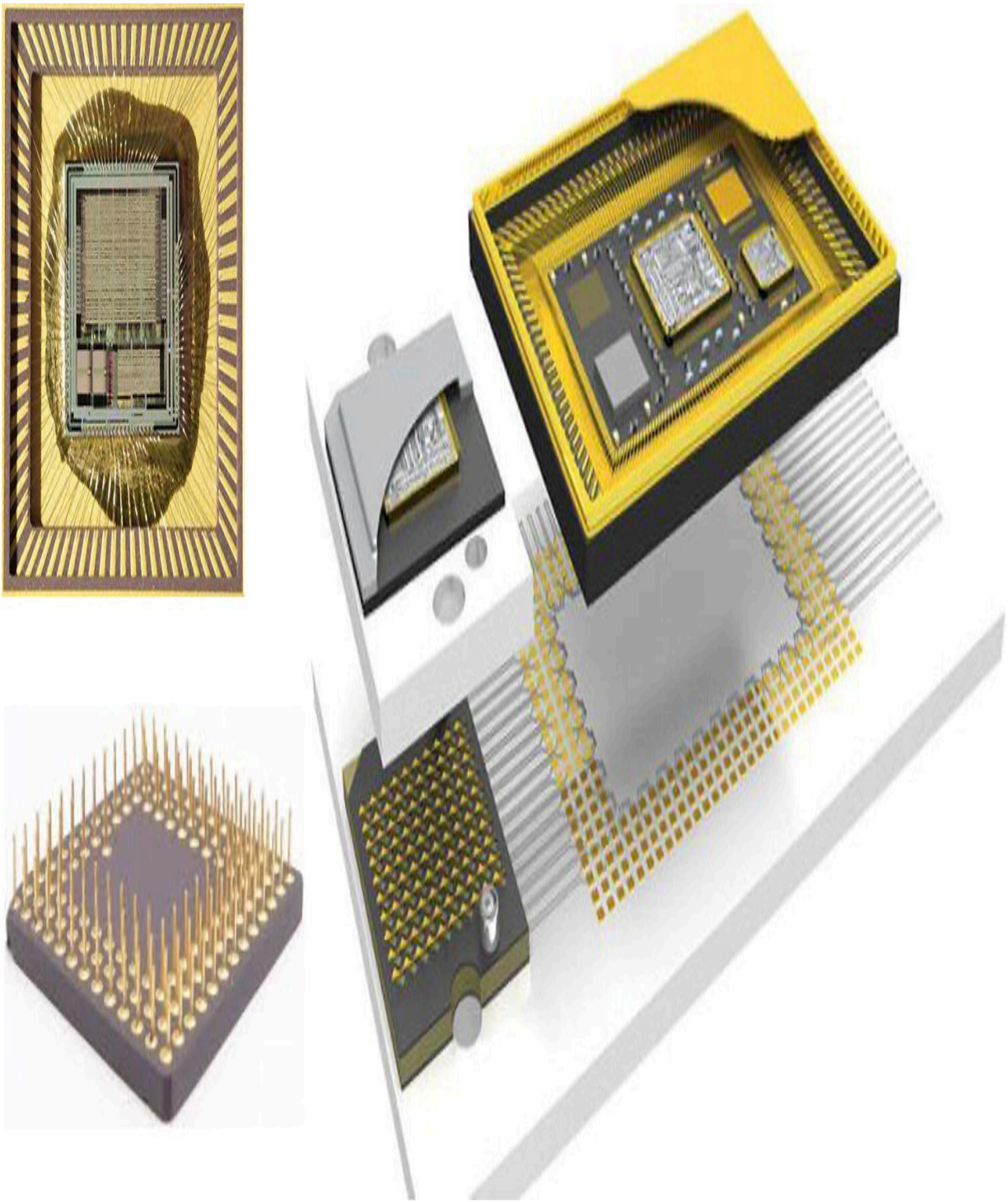


Figure 10.28 Packaging for devices with many inputs and outputs.

Source: https://thefreenewsman.com/global-ic-packaging-market-2018-j-devices-powertech-technology-stats-chippac-utac-chipmos/242011/https.

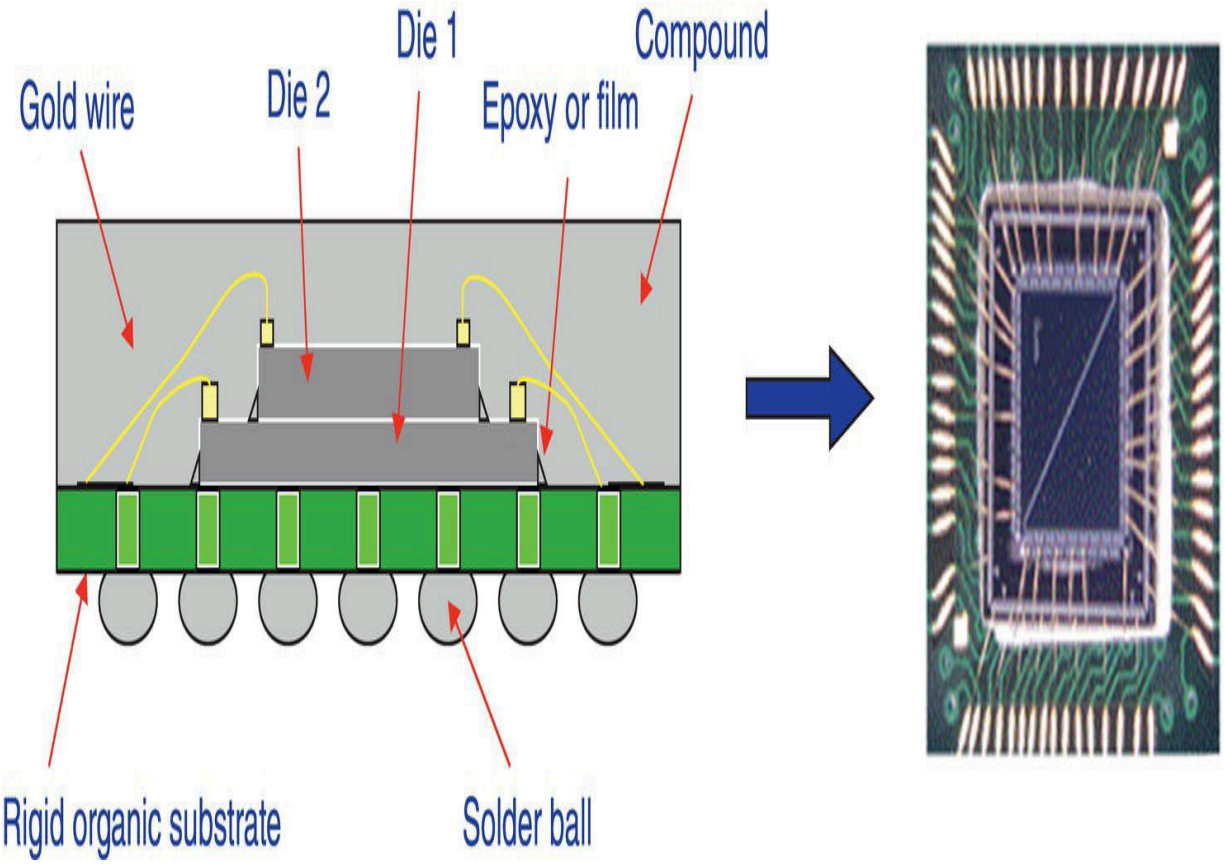


Figure 10.29 A sketch of the flip bonding process (left) and a completed packaged chip ready to be used in a system (right).

For much more complex circuits, such as processors or memories, we use packages with 100 or more contacts (Figure 10.28).

Finally, just to complete the topic of packaging, there is flip-chip bonding. We use this when we have chips with lots of inputs and outputs that are to be connected to another chip containing other type of circuits. This happens, for example, in many optical infrared devices where the detectors are made of one material (mercury-cadmium-tellurite or indium-antimonite, see Chapter 4) and the electronics are fabricated on silicon wafers. Each detector pixel has to be connected to its equivalent electronic input amplifier. Flip-chip packaging can also be used to make many connections between the silicon chip and a ceramic motherboard.

Figure 10.29 shows on the left the flip bonding concept. We have two wafers with a large number of cells, as many as a million cells in

each wafer, that have to be connected cell to cell to another chip. A diagram of these metallic bumps is shown on the right of Figure 10.29. The distance separating the bumps is just 20 μm. Usually the contacts between a silicon integrated circuit and the motherboard are much fewer than in detector technology so the pads and metal bums can be larger.

10.10 Clean Rooms

Finally, I will mention something about clean rooms. To help visualize the need for ultraclean rooms to process integrated circuits, consider that the newer processor chips are about 5 cm^2 and contain 1×10^{10} transistors (10 000 000 000) plus all the accompanying devices and aluminum connections. Simple conversion tells you that each transistor must occupy an area smaller than 5×10^{-10} cm^2. Figure 10.30 compares the size of the minimum design rules in modern processing to those of typical particles.

Minimum design rules determine the minimum distance allowed between two electrical components such as aluminum lines, contacts, p- or n- type regions or any of the many fixtures of the devices. Right now, typical minimum design rules are about 0.25 μm. Figure 10.30 compares the minimum design rules with the diameter of a human hair (300 times larger than the minimum design rules) or even a fingerprint (about 10 times larger). You can see the damage that any of these floating particles can cause if they sit on one of the wafers, or worse still if they fall on a mask. The entire lot, which may take as many as 50–200 processing steps, 15–30 photolithographic plates and three months of work can be totally destroyed by a fingerprint!

Impurities and defects in processing wafers are extremely expensive. To give you an idea, a 100 mm^2 wafer with no defects costs, say, 1. If I have one defect per cm^2, the cost goes up by a factor of 3. If I have two defects the cost goes up by a factor of 10. These are relatively small size wafers. For a 200 mm^2 wafer, again assuming

the cost to be 1 if there are no defects, the cost increases by a factor of 10 if we have one defect per cm^2 and by a factor of 25 if we have two defects per cm^2. We work now with 300 mm^2 wafers so you can see how important is to keep laboratories ultra-super clean.

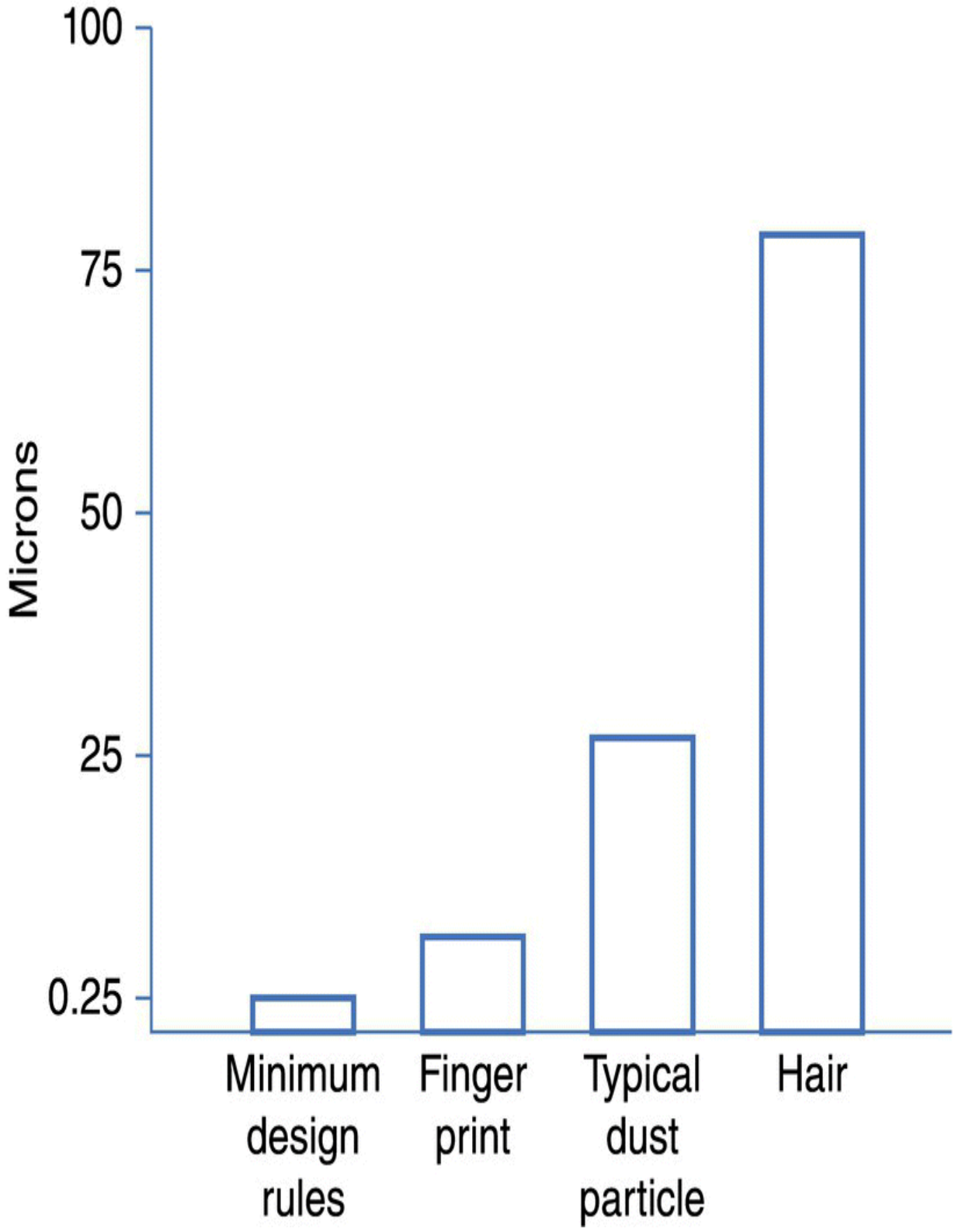


Figure 10.30 The minimum design rules compared to typical impurities that can pollute clean rooms.

The process yield is the ratio of good dice divided by the total number of dice. Take a look at Figure 10.31. All three wafers have the same number of defects, that is, five, and they are located in the same place. The chips on each of the wafers have different sizes. On the left we have small chips. I count 196 chips and 5 of them are damaged. The yield of the wafer is (196 – 5)/196 or 97.5%. If the chip is four times larger, the yield is 41/49 or 90%. If I make the chip larger still, the situation on the right of Figure 10.31, the yield now goes down to 11/16 or 69%, that is, I have to discard about one-third of the dice I fabricate.

To keep these clean rooms "electronically" clean we use laminar flow systems, as I show in Figure 10.32. After the air goes through many filters it then flows uniformly down, avoiding as many turbulences as possible (people moving, tables, and other objects can create a lot of turbulence). Additionally, workstations have their own clean-blowing systems. The air flows from the top of the station and out of the front of the table, not allowing anything floating in the air to go inside the station. The air goes through the floor, passes several fancy filters, and is then recirculated.

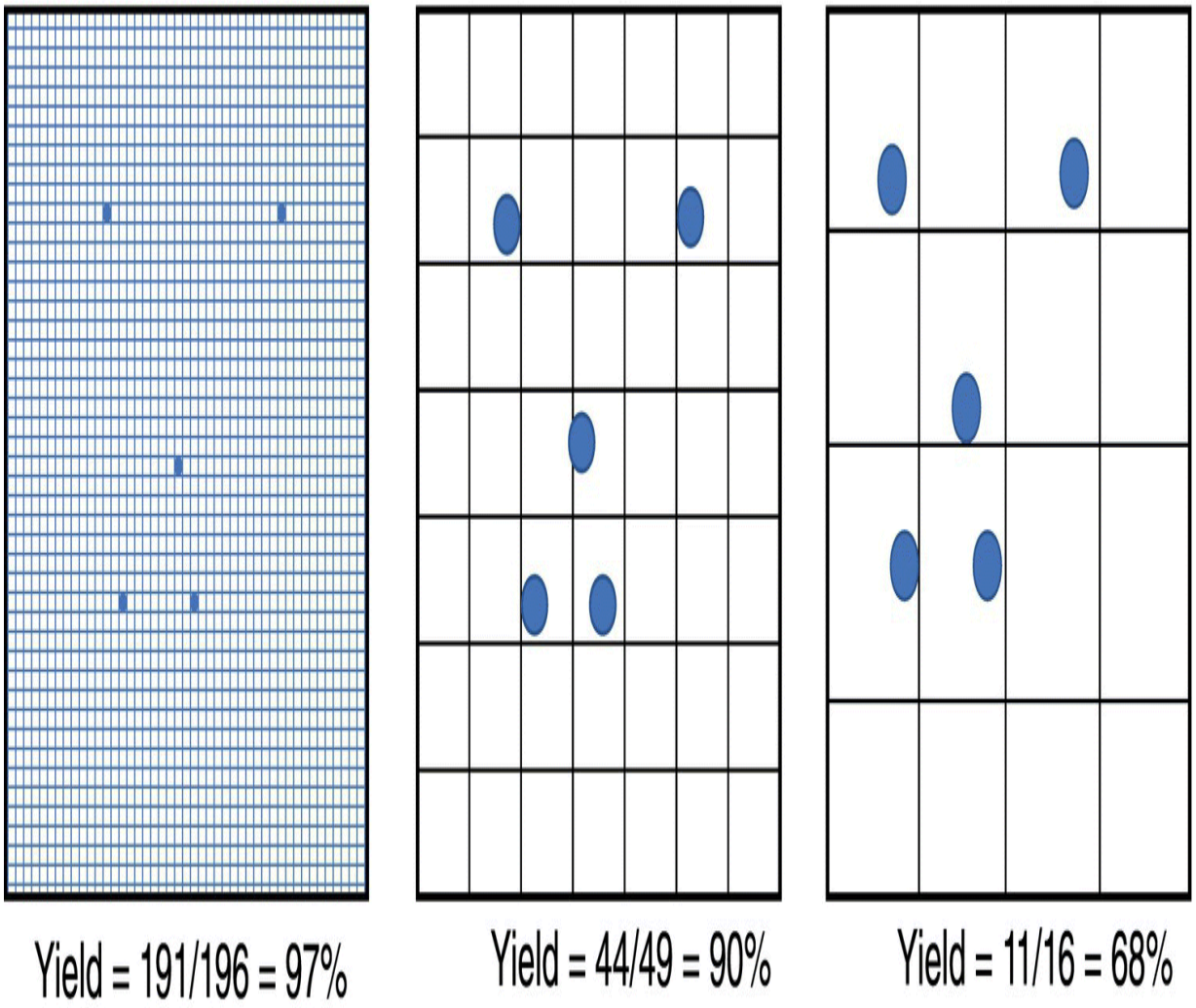


Figure 10.31 Effect on yield of defects as a function of chip size.

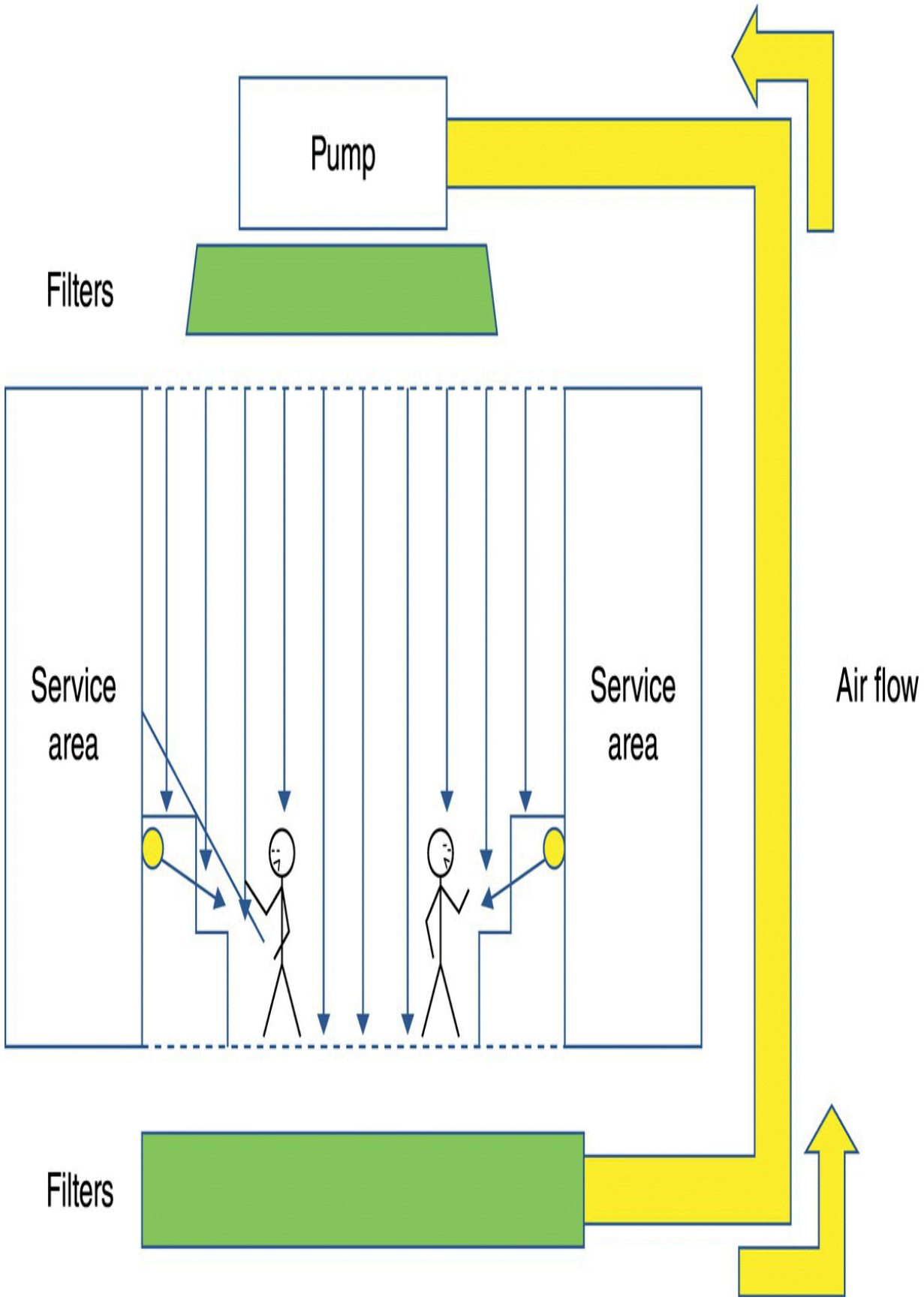

Pump
Filters
Service area
Service area
Air flow
Filters

Figure 10.32 A typical laminar flow clean room keeps the air flow vertically down, filtered, and recycled. Working areas have their own additional filtering and air circulation away from the table. Service areas are separated from the clean room.

On the sides and out of the of the clean laboratory there are service area halls. These allow the servicing of machines and any required repairs and the maintenance to be done outside the cleaning areas.

In the same way that the rooms are extraordinary clean, the personal entering the laboratory must dress in completely clean white clothes, head to toe, and go through a series of doors with flushes of air, like an air shower, to ensure that they do not bring any particles into the laboratory. (I have been inside these laboratories a few times and I admire the technical personnel that work inside them for hours. It is not a comfortable environment.) All the personnel entering the laboratory must wire themselves to ground before they touch anything to avoid any static electricity, another killer of ultra-sensitive integrated circuit chips. Much of the process is done automatically by machines, reducing the number of human beings that have to be inside the laboratories.

There is an excellent 10-minute video on YouTube from Global Foundries, an integrated circuit fabrication company, that gives you a tour of their immense clean room and goes over many of the steps I cover in this chapter. Go to https://www.youtube.com/watch?v=qm67wbB5GmI. I am sure you will enjoy it. Another YouTube video worth watching is one from Semiconductor Technology at TSMC, which is very informative (https://www.youtube.com/watch?v=4Q_n4vdyZzc).

10.11 Additional Thoughts About Processing

What I have explained above is the basic technology used in semiconductor processing today, but as you may suspect the processing equipment currently used is a little more complicated (huge understatement) that what I have described.

The photolithographic process I explained in Section 10.4 and use repeatedly in Sections 10.5 and 10.7 is accurate but very rudimentary. That is the way I did it in the early 1970s. The process is still theoretically the same, but you can imagine that the actual technology has improved several thousand times over.

We call the photographic plate a reticle. The word "mask" applies when the pattern occupies the entire wafer. The reticle contains the pattern of a single chip or die. We use reticles in the stepper projection system to repeat the same pattern over and over until the same chip is replicated on the entire wafer. The reticle is created by a computer-controlled electron beam. Figure 10.33 shows a photograph of an advanced photolithography stepper (from ASM Lithographic Co.) and one of the many sophisticated optical systems (from Corning Troper Co.).

We use lasers for higher resolution. In the 1990s we were able to get resolutions of maybe half a micron. Now we can fabricate circuits with design rules close to 10 nm, about 50 times higher resolution than 30 years ago. Some of these advances, for example photolithographic systems with special lens systems, repetitive exposure, wafer handling, and fancy light sources, can cost over a billion dollars. Some very large processing centers can fabricate up to 30 000 wafers a month.

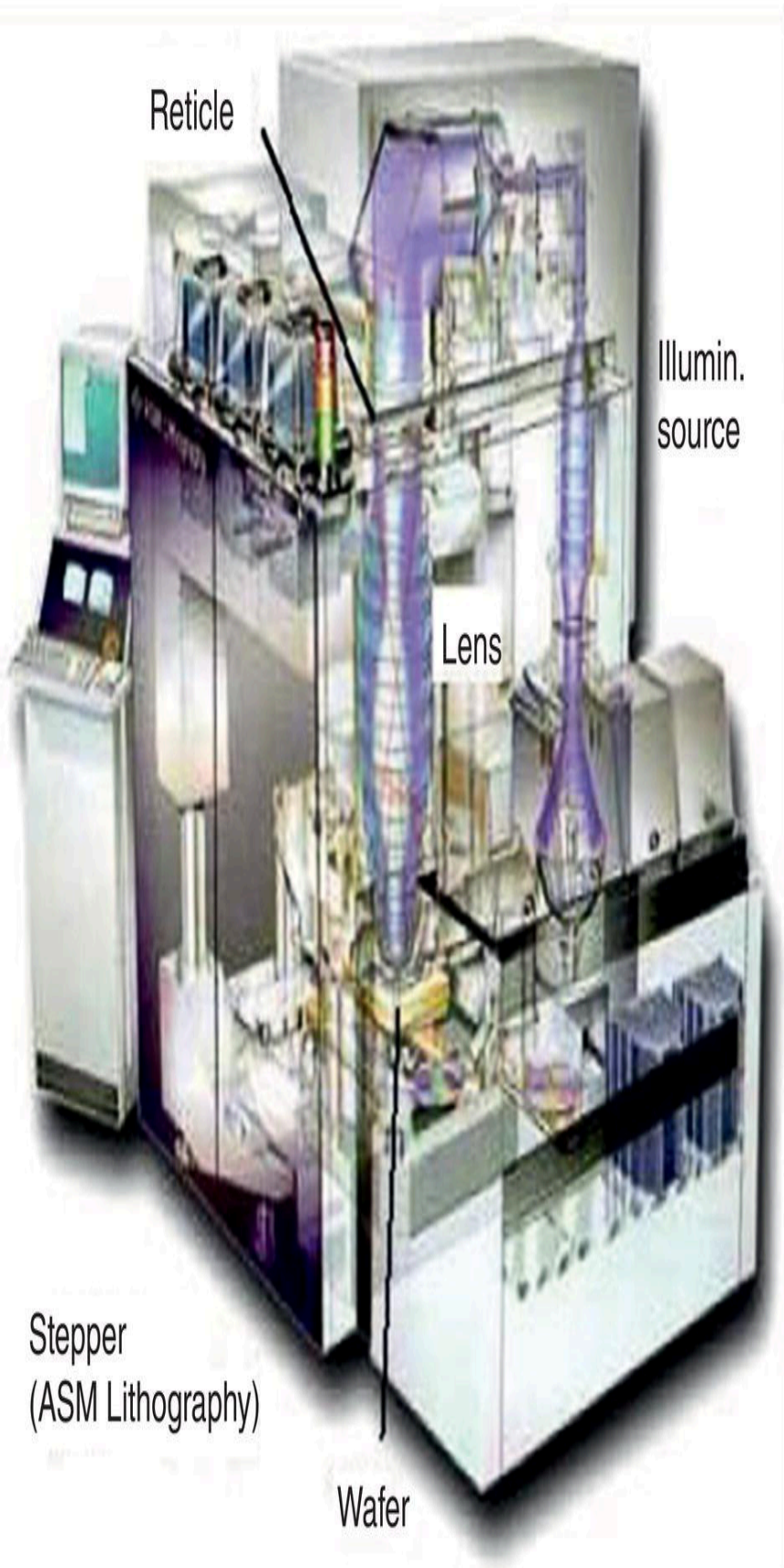

Reticle
Illumin.
source
Lens
Stepper
(ASM Lithography)
Wafer

Figure 10.33 Left, a stepper photolithography system (ASM Lithography Co.). Right, a sample of an optical lens system (Corning Traper Co.).

One of the most repeated and popular laws of electronics is Moore's law. Gordon Moore stated in 1965 that the number of transistors per square inch (6.5 cm^2) would double every year. That has been true at least until today. Figure 10.34 shows the process improvements from 1970 to 2016, going from about 2000 transistors per square inch to over 10 billion in the same space, 1 $in.^2$. By the time you read this sentence, the graph in Figure 10.34 will be already obsolete. Go to Wikipedia and look at the new numbers. They will be more impressive!

10.12 Summary and Conclusions

In this chapter I have described the process used to fabricate integrated circuits, from pure silicon to the fabrication of perfect circular boules, diced and marked to generate uniform wafers. We have seen that this can be done by growing a boule from a melt, the Czochralski method, or by purifying it using the flat-zone method. We also saw that for additional purity and regularity we can add by epitaxial methods a thin film of much more controlled material, an epitaxial layer.

By repeatedly growing oxide on the surface of the semiconductor, covering the oxide with photoresist, exposing the photoresist to light, removing the exposed area of the photoresist, and removing the oxide under it, we can create a wafer with protected and unprotected regions. We can then add impurities in these isolated areas by thermal diffusion or implant different impurities on them to form the p- and the n-regions of the semiconductor. A similar procedure allows us to deposit the metals that connect the different semiconductor components. Finally, I discuss very briefly the equipment needed to make these devices.

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.

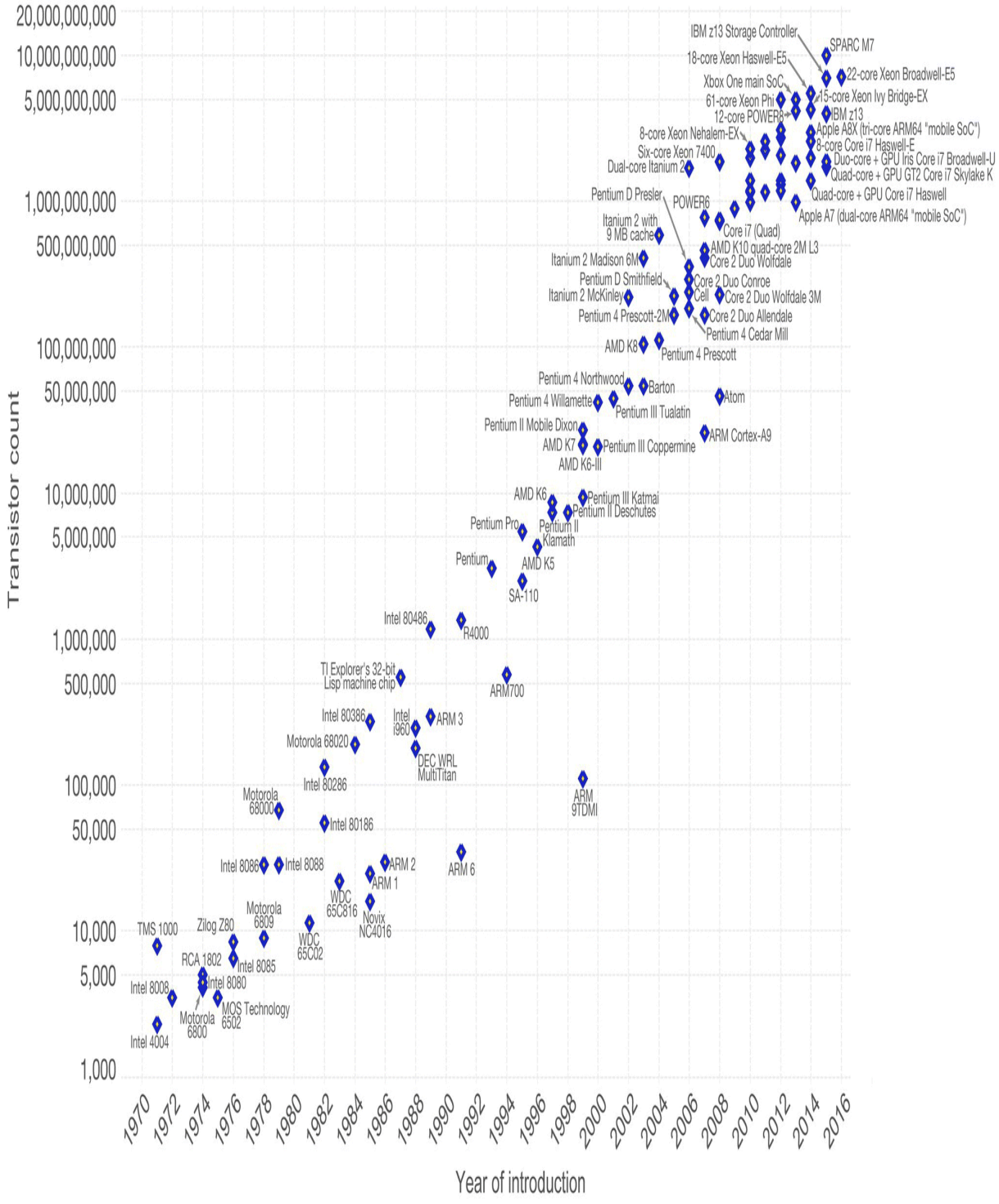


Figure 10.34 1970 to 2016 progress in the transistor count per square inch.

Source: https://en.wikipedia.org/wiki/Transistor_count#/media/File:Moore's_Law_Transistor_Count_1971-2018.png; https://en.wikipedia.org/wiki/Moore%27s_law.

Appendix 10.1 Miller Indices in the Diamond Structure

At the end of Section 10.2.1 I mentioned that when we finish growing the boule, we grind it to make a perfect cylinder. This makes sense since we want all the wafers to have exactly the same diameter. Then I mentioned that we make a notch on one side of the boule so that when we slide the boule, we know which is the crystallographic direction of the wafers. Why do we do this?

Let me introduce the concept of Miller indices. Figure 10.35 shows the crystal structure of diamond, the same drawing as in Figure 3.1, and three ways in which I can slice the wafers. In the middle of Figure 10.35 I show three ways I can slice a diamond lattice crystal structure. At the top of Figure 10.35, the cut designated as (1,0,0), I slice the crystal along one face of the cube, shown as a shaded area. If you compare the (1,0,0) cut to the crystal structure on the left, you will notice that this cut contains four atoms at the four corners and one more in the middle (the ball with the two chopped bonds). Only one quarter of the corner atoms belongs to each unit area, but the atom in the middle, the blue atom in the structure, is inside the box so the cut does not slice that atom. Thus, each unit area in the (1,0,0) cut includes two atoms:

$$\text{number of equivalent atoms in the (1,0,0) cut} = \left(4 \times \frac{1}{4}\right) + 1 = 2 \tag{10.1}$$

If the atomic distance between atoms is *a*, then the unit area for this slice is a^2.

The diagonal cut in the middle, the (1,1,0) orientation, again looking at the crystal structure, goes from one corner to the opposite one, cutting the atom in the middle and contains a quarter of four atoms in the corners, two half atoms in the middle, and two full atoms inside. The number of equivalent atoms associated with this plane is four:

$$\text{number of equivalent atoms in the (1,1,0) cut} = \left(4 \times \frac{1}{4}\right) + \left(2 \times \frac{1}{2}\right) + 2 = 4 \tag{10.2}$$

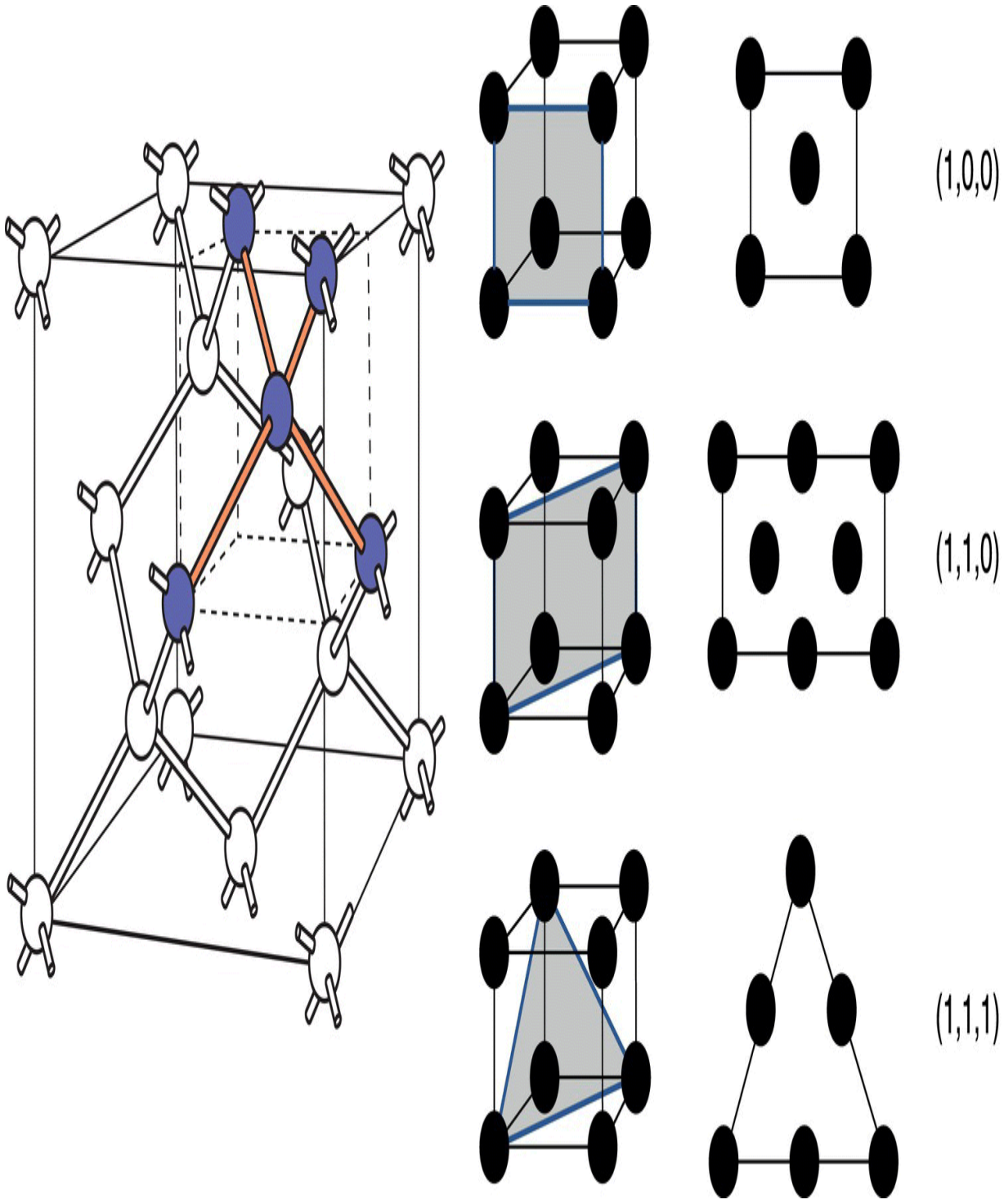


Figure 10.35 Three ways we can slice the diamond crystal structure.

However, the area is larger because the vertical side of the cut is *a*, the lattice constant, but the horizontal side of the cut is the

hypothenuse of a right-angled triangle, so its length is given by

$$\sqrt{a^2 + a^2} = \sqrt{2}a$$

and the area is

$$a \times \sqrt{2}a = \sqrt{2}a^2 = 1.414a^2 \tag{10.3}$$

Finally, the triangular cut (1,1,1) has three corner atoms and three middle atoms. The atoms in the corners now are shared by six slices and the three in the middle by two planes, so the total number of equivalent atoms per unit area is

$$\text{number of equivalent atoms in the (1,1,1) cut} = \left(3 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{2}\right) = 2 \tag{10.4}$$

The area of an equilateral triangle is $\sqrt{3}/4$ times the length of one of the sides, which in this case is $\sqrt{2}a$, thus the area of the triangle in the (1,1,1) cut is

$$\frac{\sqrt{3}}{4} \times \left(\sqrt{2}a\right)^2 = \frac{\sqrt{3}}{2}a^2 = 0.866a^2 \tag{10.5}$$

Knowing that the lattice constant for silicon is $a = 5.43$ Å $= 5.43 \times 10^{-8}$ cm, then $a^2 = 2.95 \times 10^{-15}$ and we can now calculate the atomic density of atoms for each slice and we get, in incremental order:

$$D_{100} = \frac{2\ (\text{atoms})}{a^2\ (\text{cm}^2)} = \frac{2}{2.95\times10^{-15}} = 6.77\times10^{14}\ \text{atoms-cm}^{-2}$$

$$D_{110} = \frac{4\ (\text{atoms})}{1.414a^2\ (\text{cm}^2)} = \frac{4}{1.414\times2.95\times10^{-15}} = 9.59\times10^{14}\ \text{atoms-cm}^{-2}$$

$$D_{111} = \frac{2\ (\text{atoms})}{0.866a^2\ (\text{cm}^2)} = \frac{2}{0.866\times2.95\times10^{-15}} = 7.83\times10^{14}\ \text{atoms-cm}^{-2}$$

The diagonal cut (1,1,0) has the highest density of atoms per layer and therefore this cut is preferred for hole conduction. For electron conduction, the straight cut (1,0,0) is preferred because it has the smallest number of incomplete bonds and the fewest dangling bonds. At every surface there are dangling bonds. The silicon atoms at the surface use two or three electrons to bond with the other atoms, but there is nothing to bond outside the solid, resulting in the dangling bonds.

After the wafers have been cut, there is no way of knowing what type of wafer it is or what orientation it has. There is a convention so we know what wafer we are dealing with. Figure 10.36 shows the convention we use. The main flat or notch is cut along the (1,1,0) direction on all the wafers. The second cuts, which are much smaller than the one at the bottom, are at 45°, 90° or 180° from the main flat, indicating that the wafer is (1,1,1) n-type, (1,0,0) p-type or (1,0,0) n-type, respectively. If there is no second cut, the water is a (1,1,1) p-type. For the larger wafers, we use a notch instead of a flat. The flat loses valuable area as the wafers get larger. We need to know the orientation because some processes can be anisotropic, that is, they have different properties when they are used in different orientations.

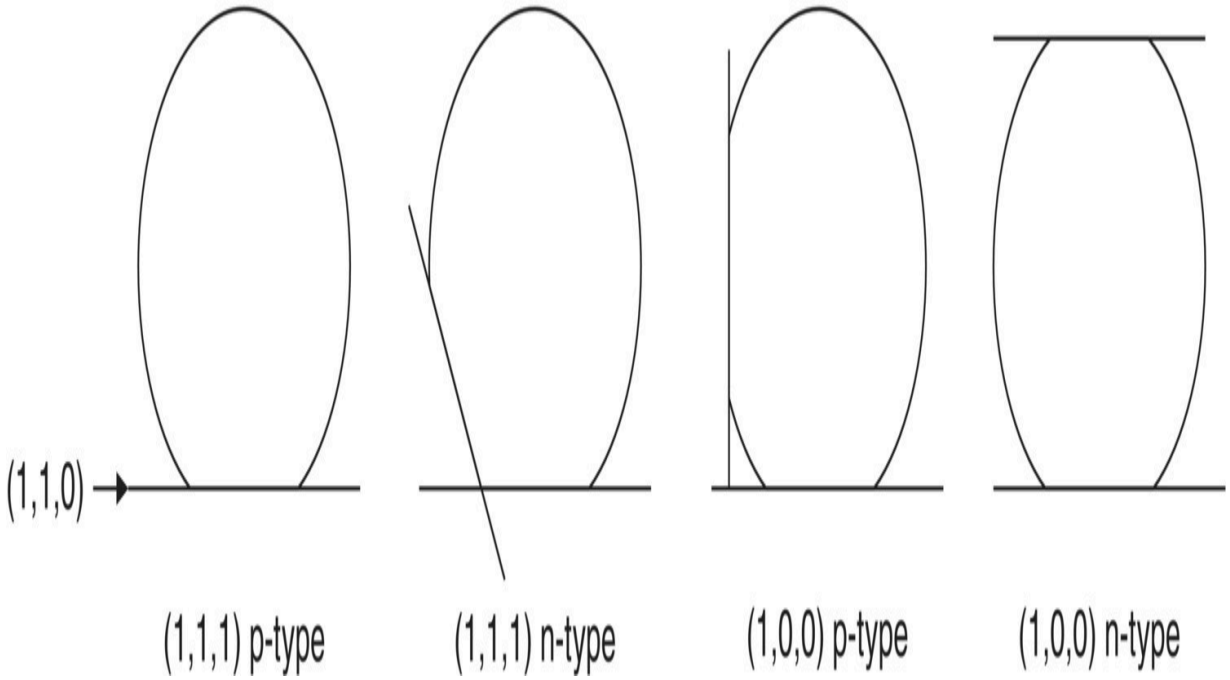


Figure 10.36 The flats in different locations around the periphery of the wafer indicate the type of wafer and its orientation.

11
Logic Circuits

OBJECTIVES OF THIS CHAPTER

Now we are going back to talk about how we use semiconductor devices to perform useful operations. In the first two sections we'll talk about the way we interphase with the computer and the basic language we use, that is, Boolean algebra and logic symbols. Then I will explain how we implement this algebra with switches (to clear the concepts) and semiconductor devices, and we'll see how to do arithmetic operations, sums, subtractions, multiplications, and divisions with the devices we covered in the previous chapters.

In the previous chapters I very much concentrated on the performance of individual components and how to use them, mainly in the analogue mode. Now we are going digital.

11.1 Boolean Algebra

Everybody knows that digital computers work with 1s and 0s. Computers do not understand the number 3. All the computations that computers have to carry out to give us any meaningful results are done using the ON and OFF conditions, ON for 1 and OFF for 0. The large TV screen that give us beautiful and sharp pictures with bright colors is based on millions of points that can be ON or OFF. Each point of light consists of three miniscule LEDs of three different colors. How a computer manipulates all this data is based on the concepts of Boolean algebra. (Forget about the word algebra. There are no equations involved outside of adding and subtracting.)

Mr. George Boole (1815–1864; Figure 11.1) was a British mathematician and philosopher with interest in strengthening the logic concepts. Aristotle is credited with creating logic thinking with his famous syllogisms, such as “All men are mortal, I am a man, therefore I am mortal” or, more abstractly, “All A are B, all B are C, therefore all A are C”. He wanted to be sure that people were logical and consistent in advancing any idea. What Boole did was to add mathematical formalism, symbolic logic, to Aristotle's logic. Now Boolean logic or Boolean algebra is the basis of all digital computer calculations. Boolean logic consists basically of three operations: AND, OR, and NOT (actually he also had the operation IMPLY but we do not use it in computers). Let's take a look at these operations. First, I will use relays to explain how the three operations work and then we'll step up to using CMOS.

Figure 11.1 George Boole developed the symbolic logic language called Boolean algebra.

Source: https://en.wikipedia.org/wiki/George Boole#/media/File:George_Boole_color.jpg.

11.2 Logic Symbols and Relay Circuits

A relay is a simple electromechanical switch that works in the same way as a remote switch. You press a button, a signal goes to the relay, a current goes through its coil and this acts as a magnet and turns the switch ON. Figure 11.2 shows sketches of two relays. The one on the left is a normally OFF relay, that is, the switch is normally open, and it closes, turns ON, when I apply a voltage. The relay on the right is a normally ON, that is, the relay opens, turns OFF, when we apply a voltage to the coil.

Now let us see how we can use these relays to make a logic circuit. The diagram at the top left of Figure 11.3 shows the circuit for the logic operation AND. The relays can have two values, ON and OFF, but the light will turn ON only if both relays A and B are ON. These conditions are shown in the truth table on the right of Figure 11.3. We call these tables truth tables because they show the true output based on the state of the two inputs (it is true that when one switch is OFF and the other is ON the output, the light, is OFF). The name truth table came from the logicians whose interest was to logically know what the true result of a logical thought is. In electronics we use the symbol at the bottom left of Figure 11.3 for AND operations. When we use this symbol, we do not care what is inside the symbol (could be relays, transistors or a trained monkey), just that the relation between the two inputs, A and B, results in an output, C, that satisfies the conditions of the AND function.

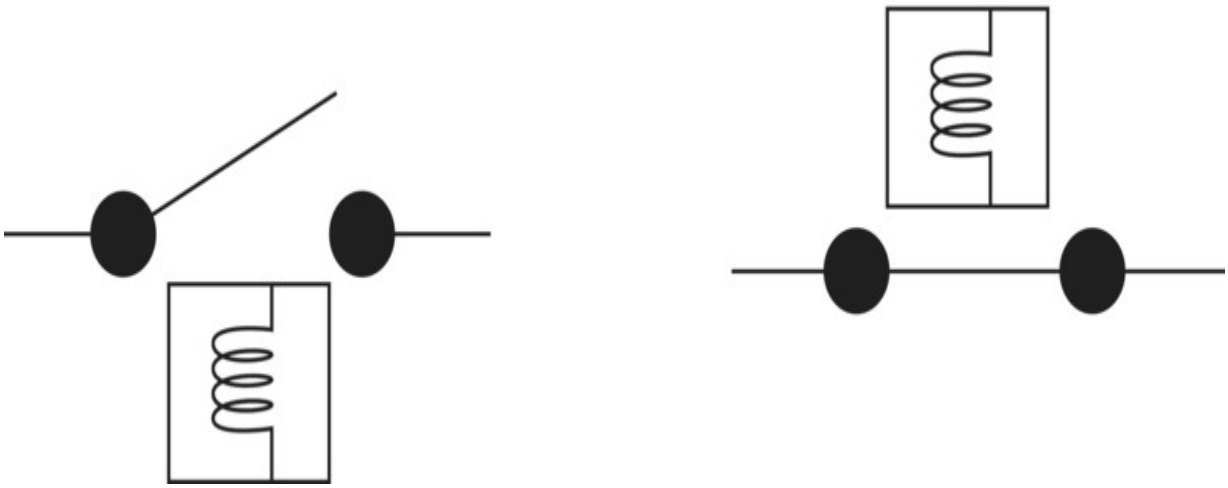
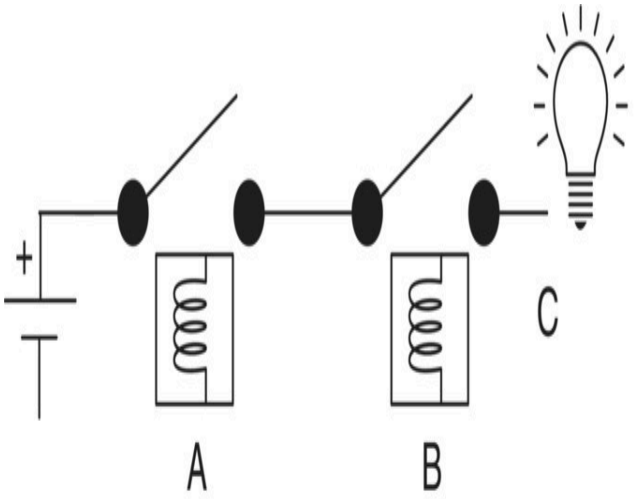


Figure 11.2 Symbols of normally OFF (left) and normally ON (right) relays.



Switch A	Switch B	Output, C
OFF	OFF	OFF
ON	OFF	OFF
OFF	ON	OFF
ON	ON	ON

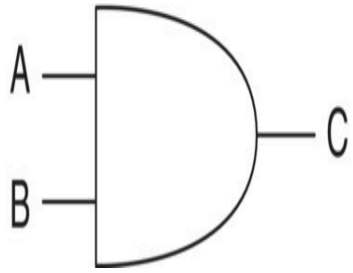
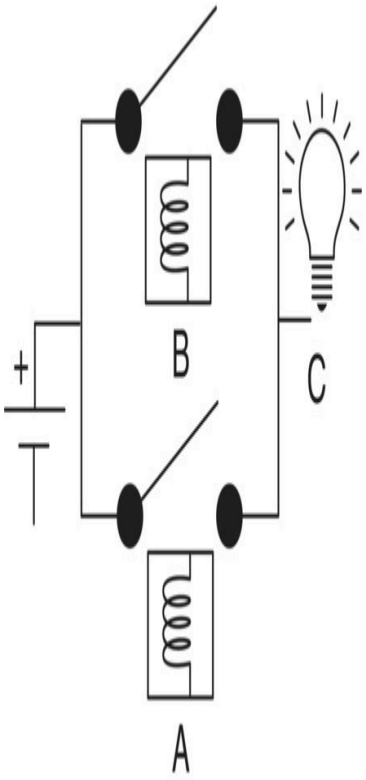


Figure 11.3 The logic circuit AND using two normally closed relays (top left), the symbol we use for the AND operation (lower left), and the truth table (right). Both switches have to be ON for the light to turn ON.



Switch A	Switch B	Output, C
OFF	OFF	OFF
ON	OFF	ON
OFF	ON	ON
ON	ON	ON

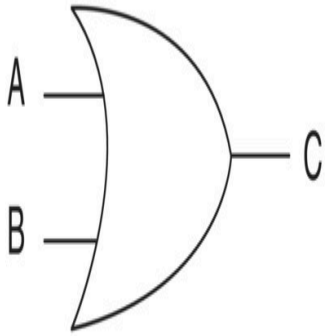
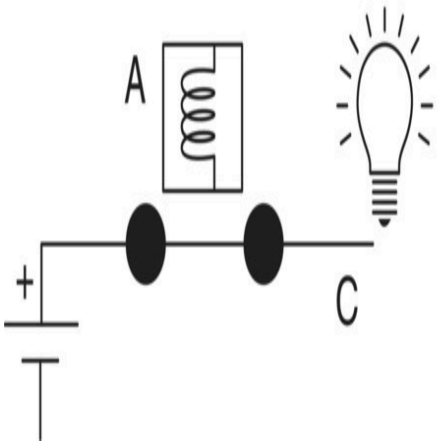


Figure 11.4 The logic circuit OR using relays (left), its truth table (middle), and the symbol for OR (right). If either of the two relays is ON, the light will be ON.



Switch A	Output C
OFF	ON
ON	OFF

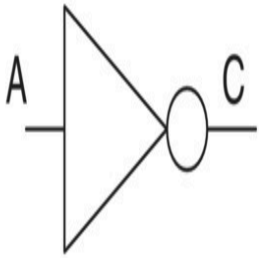


Figure 11.5 The logic circuit NOT using a relay (left), the truth table (middle), and the NOT symbol (right). We use a normally closed relay, so when the relay is OFF the light is ON and vice versa.

Figure 11.4 shows the OR function. Now the light is ON if either switch A or B or both are ON. On the right is the symbol we use for the OR function and in the middle is the truth table for the OR function. Finally, Figure 11.5 shows the circuit, the truth table and the symbol for the NOT operation, also called the inversion operation.

These three circuits, AND, OR, and NOT, are the main Boolean operations that we use in electronic designs. For the design of digital functions, and for pure convenience, we like to use four additional operations. The first one is called the exclusive OR or XOR operation. I show it in Figure 11.6. I do not show the equivalent relay network for the XOR circuit because it would be more confusing than helpful. But you can see that, by definition, the XOR is ON if and only if just one, and only one, of the inputs is ON. If both are ON or both are OFF, the output is OFF.

Switch A	Switch B	Output C
OFF	OFF	OFF
ON	OFF	ON
OFF	ON	ON
ON	ON	OFF

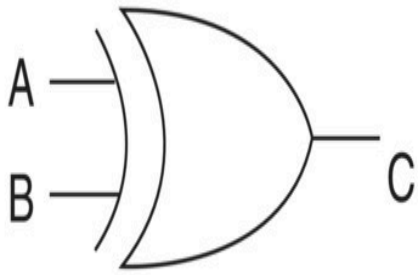


Figure 11.6 The XOR truth table (left) and its symbol (right). For the output to be ON, one and only one of the inputs has to be ON. The output is OFF when both inputs are ON or both are OFF.

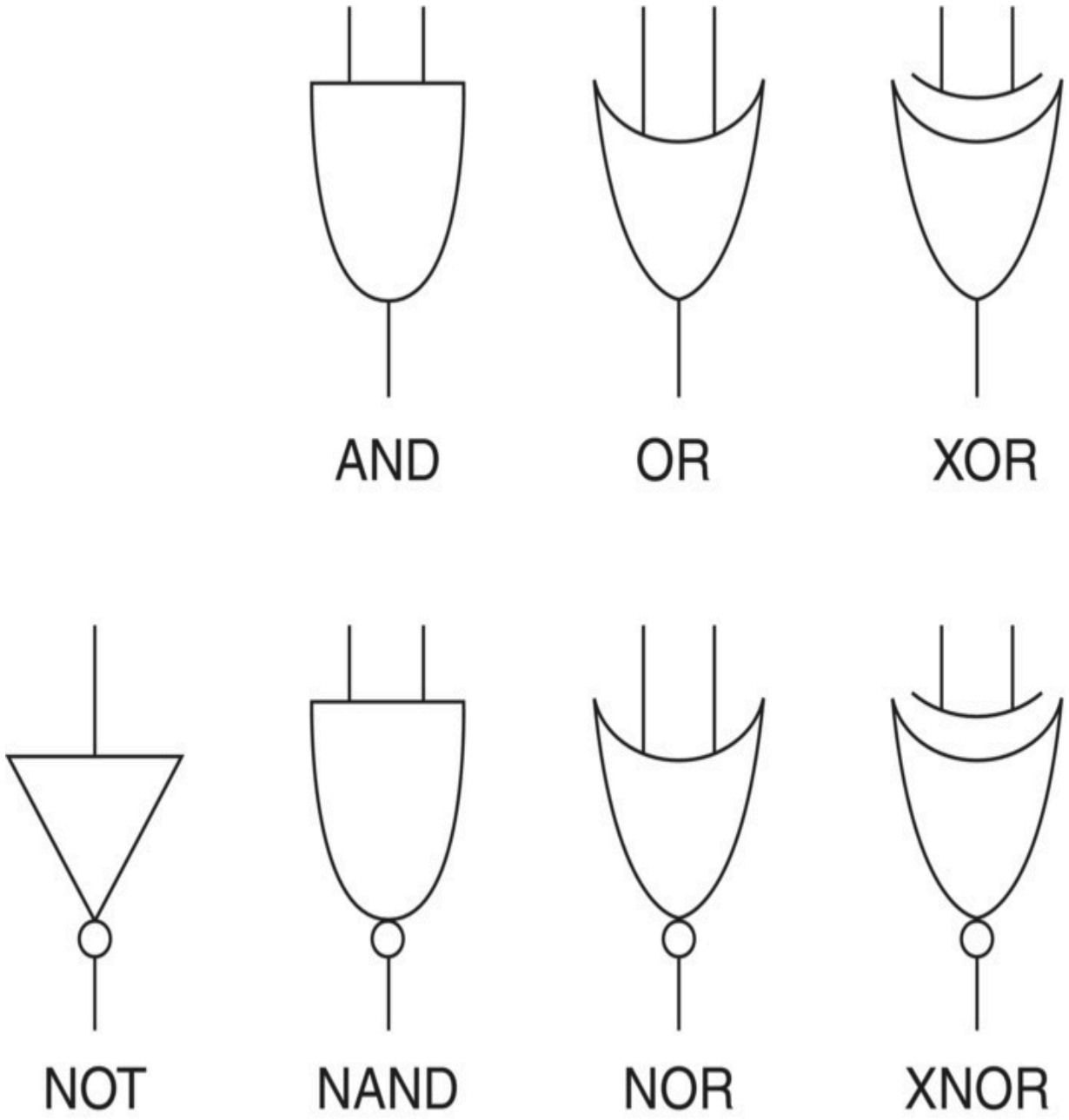


Figure 11.7 The seven logic symbols we use in designing digital electronic circuits.

These symbols are the basic components used in electronics and from them we can define others that we'll use in subsequent sections. The added ones are basically the reversal of the first three. Figure 11.7 shows all the logic symbols that we use. It is actually quite simple to remember these symbols and memorizing them will help you to follow the circuits that I explain later on. The only

symbol with a triangle is the NOT circuit. If the line on top of the ovals is straight, the symbol is the AND function, if curved, it is the OR function, and if it has two curves on top, it is the XOR function. These last three may have a small ball at the output terminal, as I show in the second row of Figure 11.7. This ball indicates that the truth table is exactly the opposite of the one above without the ball, that is, the NAND circuit is always ON except when the two inputs are ON, the opposite of AND. The same for the other two circuits. We added these "negation" circuits for convenience. We can get any one of them by adding a NOT circuit to any of the other "positive" symbols.

11.3 The Electronics Inside the Symbols

We do not use relays in electronic circuits (although in the 1940s and 1950s some engineers, including myself, designed logic machines [computers?] using relays, see Appendix 11.5) so next I discuss how to create these truth tables using semiconductor devices, first using diodes and then using MOSFETs (Metal-Oxide-Semiconductor Field-Effect-Transistor). For MOSFETs I use the terminology CMOS (Complementary MOS), which stands for complementary MOS. It is a way of indicating that I use both p- and n-type MOS in many circuit constructions.

11.3.1 Diode Implementation

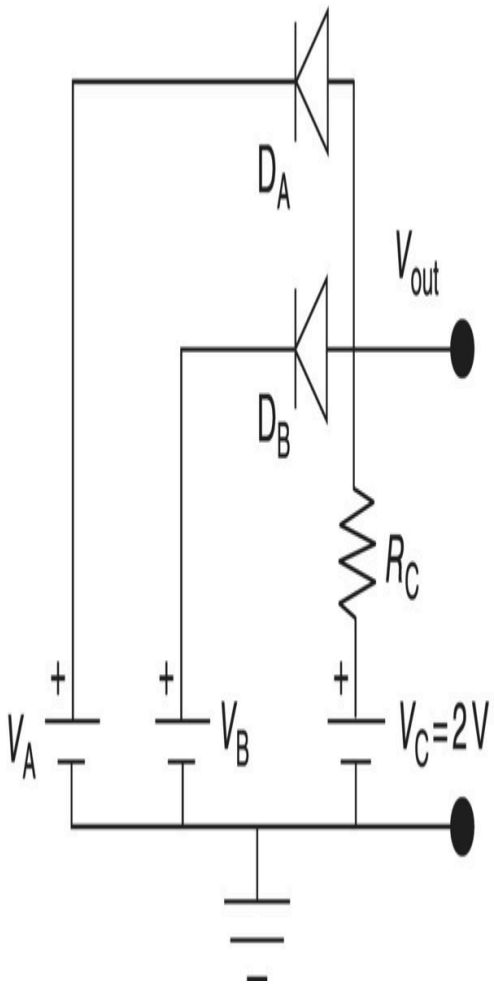
You may wonder what is inside the symbols I show in Figure 11.7. Since the purpose of this book is to give you an idea of how semiconductor devices work, let me explain the electronics inside these logic modules. First, let's use just diodes. Figure 11.8 shows the implementation of the AND function using diodes. In this and all following truth tables I will use 1 and 0 instead of ON and OFF.

In this implementation, we assume that both diodes are ideal, that is, when they are reversed biased, the diodes are OFF and their resistance is infinite. When they are forward biased the diodes are ON and their resistance is zero. I also assume that the output is

connected to a high input resistance device so the output resistance does not affect the operation of the AND circuit.

If both input voltages, V_A and V_B, are ON, equal to 1 V, the output voltage, V_{out}, has to be 1 V. Why? If V_{out} were larger than 1 V, let us say 1.5 V, the diodes would be forward biased. But a forward biased diode has zero voltage across it, so V_{out} has to be equal to V_A or V_B. and there will be a 1 V voltage drop across the resistor R_C. If the output were to be less than 1 V, let us say 0.5 V, then the diodes would be reversed biased, that is OPEN and there would be no current through R_C and therefore no voltage across it, and V_{out} would have to be equal to V_C, which is equal to 2 V, but that would make the diodes forward biased which would force the output again to be equal to V_A or V_B, which contradicts the assumption that V_{out} is less than 1.

Now suppose either V_A or V_B or both is zero. Then one or both diodes are forward biased, shorted to ground, and the output is zero. This satisfies the truth table shown in Figure 11.8.



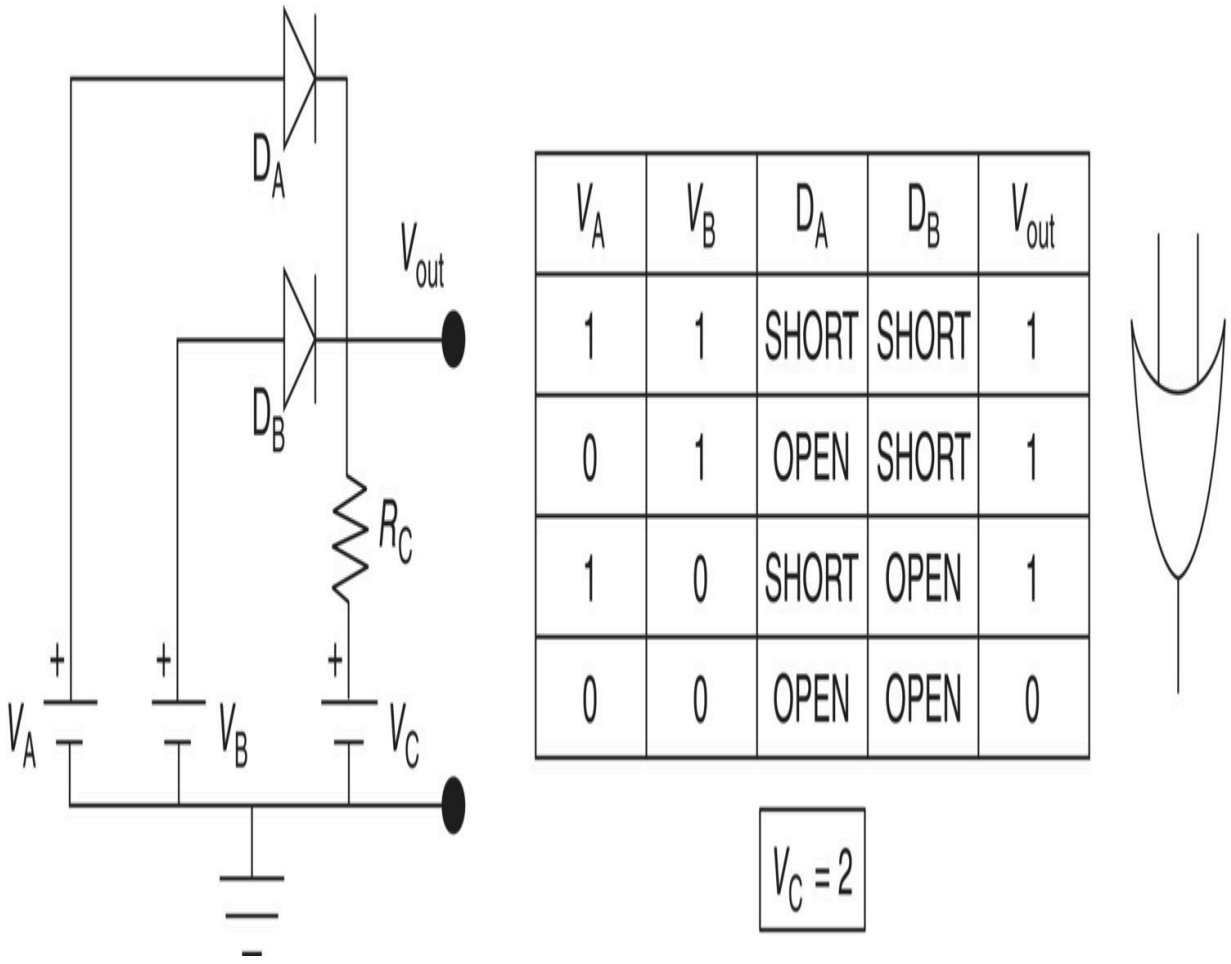
V_A	V_B	D_A	D_B	V_{out}
1	1	OPEN	OPEN	1
0	1	SHORT	OPEN	0
1	0	OPEN	SHORT	0
0	0	SHORT	SHORT	0

Figure 11.8 Diode implementation of the AND function (left), the truth table (middle), and the symbol (right). There is current through the resistor R_c only if both V_A and V_B are ON.

The implementation of an OR circuit using ideal diodes is easier to follow (Figure 11.9).

Notice that the OR circuit is the same as the AND circuit but with the two diodes reversed. In this case, if either V_A or V_B or both is 1, then one of the diodes, or both, is forward biased, that is, shorted, and the output voltage is 1, creating the conditions I show in the truth table. If V_A is 1 and V_B is zero, then the diode D_A is forward biased and therefore shorted, but diode D_B is reversed biased thus open, preventing the current from flowing to ground. Only when

both V_A and V_B are both zero will the output be zero. This defines the operation of an OR circuit.



V_A	V_B	D_A	D_B	V_{out}
1	1	SHORT	SHORT	1
0	1	OPEN	SHORT	1
1	0	SHORT	OPEN	1
0	0	OPEN	OPEN	0

$V_C = 2$

Figure 11.9 Diode implementation of an OR function (right) with the truth table (middle) and the symbol (left). The output voltage will be zero only when both inputs are zero.

11.3.2 CMOS Implementation

I will use the symbols shown in Figure 11.10 for the CMOS. I use an open circle at the contact with the base to indicate that the MOSFET is a p-type or no circle if it is an n-type. You can see in the sketches where the gates of the CMOS are, but notice that I do not tell you which terminal is the source and which is the drain. CMOS are bidirectional, so which terminal is which depends on how I connect them and in which direction the current flows.

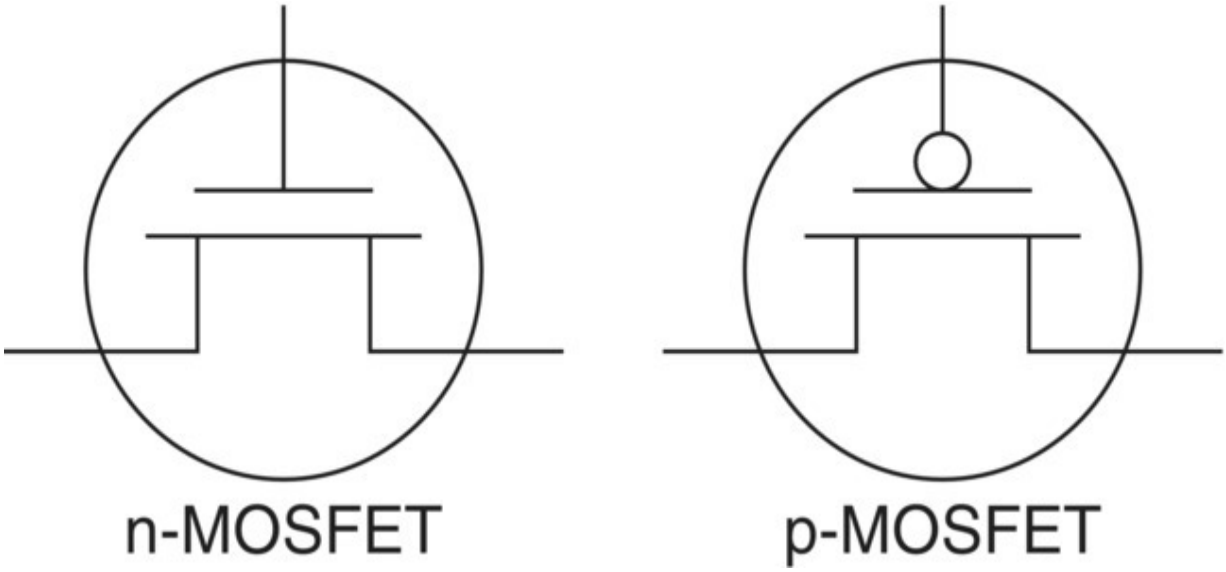


Figure 11.10 Symbols for the n- (left) and p- (right) MOSFETs. The p-MOSFET has a dot at the gate.

11.4 The Inverter or NOT Circuit

Figure 11.11 shows the inverter of the OR circuit, the NOT circuit, with the truth table and its symbol. I use both p- and n-type MOSFETs. This is why we use the term CMOS.

Notice that, when V_{in} is OFF, 0, the n-type MOSFET is turned OFF and the p-type MOSFET is ON. I represent this condition by replacing the n-type MOSFET by an open circuit and the p-type MOSFET by a short circuit. The ON MOSFET has very little resistance so the output voltage is very close to voltage V_{CC}. The output is therefore ON, as the truth table shows in Figure 11.11.

When the input voltage, V_{in}, is 1, that is, ON, the opposite occurs: the p-type MOS is OFF, open, and the n-type MOS is ON, shorted. Now the output is shorted through the ON n-type MOS to ground, forcing $V_{out} = 0$, confirming that the circuit inverts the input voltage. When the input voltage is high, the output voltage is low and vice versa. That is an inversion or the NOT circuit. Quite simple.

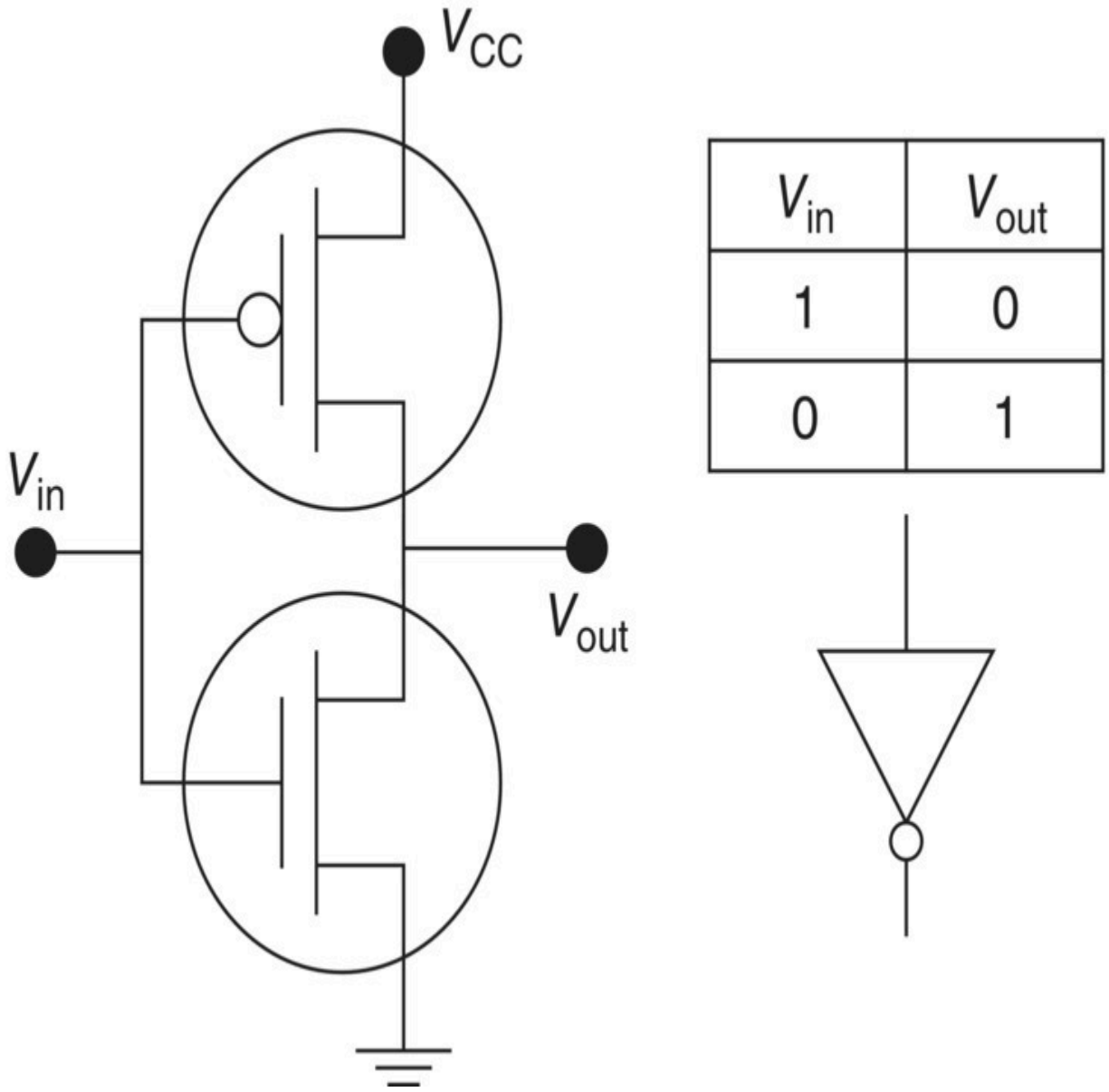


Figure 11.11 The NOT circuit using CMOS with the truth table and its symbol.

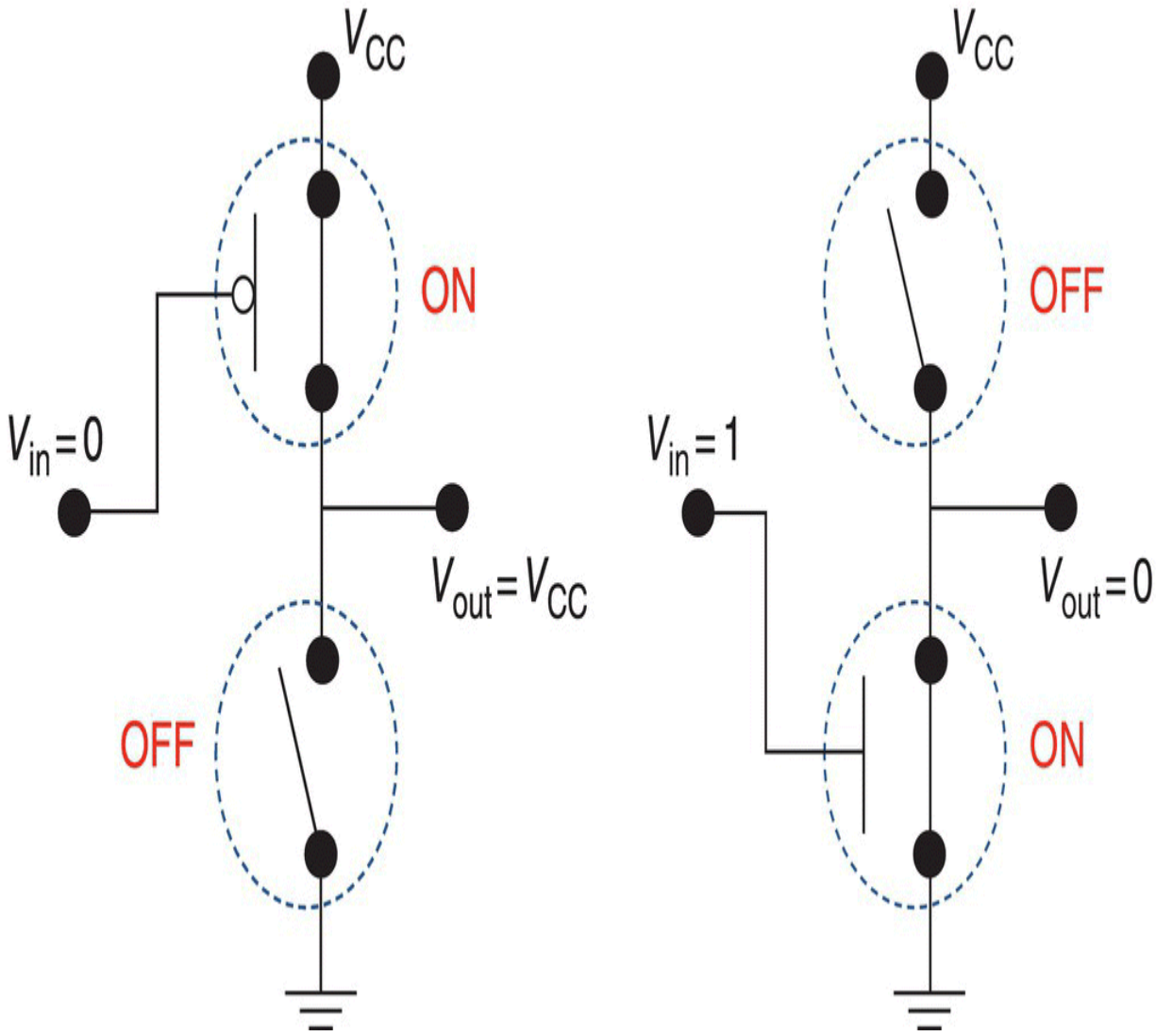


Figure 11.12 The two states of the OR circuit, with V_{in} OFF on the left and V_{in} ON on the right.

I should say that this circuit, as well as all the others I discuss next, can be implemented in much more sophisticated ways. I am just showing the simplest way, so you can get an idea of how these logic circuits are created and how they work.

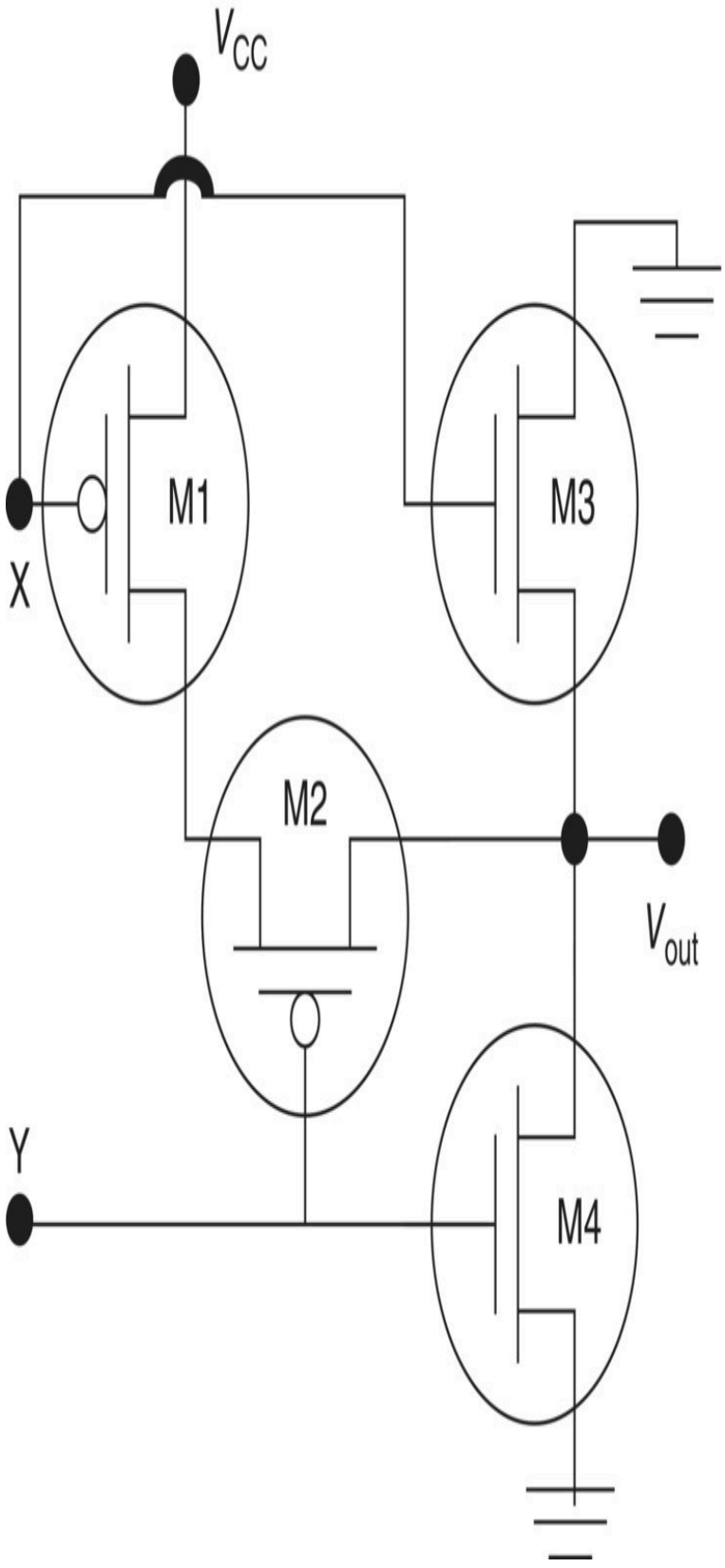
11.5 The NOR Circuit

You may wonder why I discussed the NOR circuit first instead of the OR. The NOR circuit is actually easier to implement than the OR and

to get from one to the other we just need to use the NOT circuit I discussed in the previous section, that is, the negative or inverted NOR is OR and vice versa.

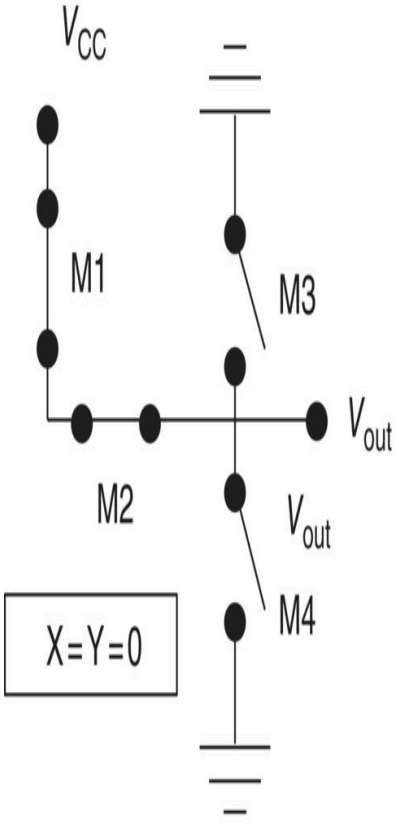
Figure 11.13 shows the NOR circuit, the truth table, and its symbol, and Figure 11.14 shows the MOSFET status for the four input combinations of 0s and 1s. I place these to figures together so you can easily go from one to the other.

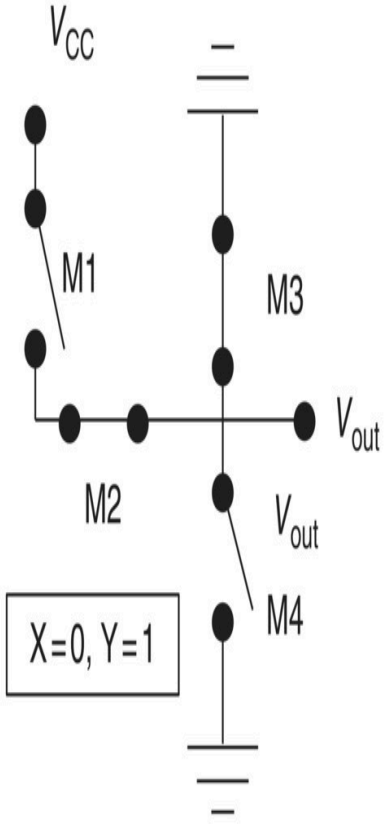
Although the figures may scare you, they are actually quite simple. Let's first take a look at Figure 11.13. It consists of four MOSFETS, two p-type (remember the ones that have the circle at the gates) and two n-type. Each input, X and Y, is connected to the gate of one p-type and one n-type MOS. The input X is connected to the gates of CMOSs M1 and M3 and the input Y to the gates of M2 and M4. These two n-CMOS are grounded on one side and joined together at the output.



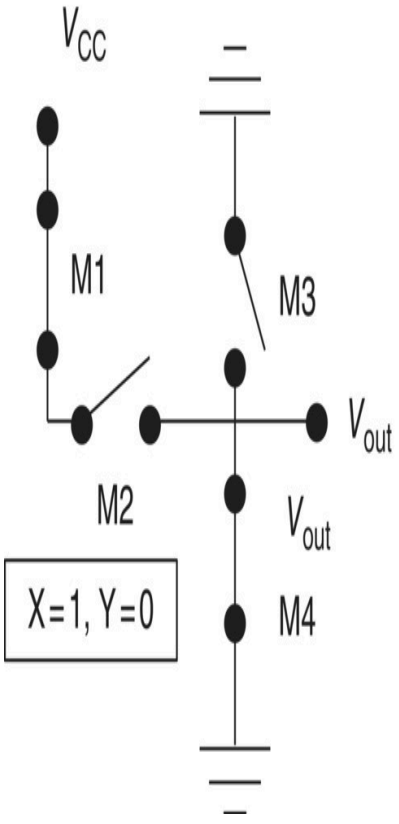
X	Y	V_{out}
0	0	V_{CC}
0	1	0
1	0	0
1	1	0

Figure 11.13 The NOR circuit (left), the truth table (top right), and the NOR symbol (bottom right).





X	Y	V_{out}
0	0	V_{CC}
0	1	0
1	0	0
1	1	0



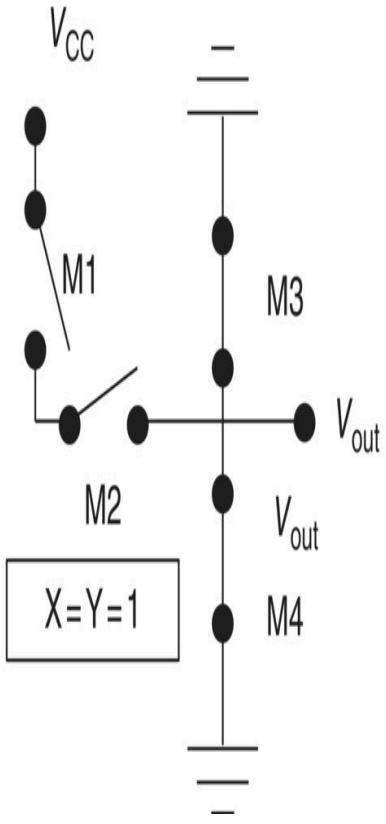


Figure 11.14 The switching status of the four MOSFET circuits as the two inputs change independently from 0 to 1. Only when both inputs are 0 is the output connected to the source, 1. In all the other three cases, the output is grounded, 0.

Now take a look at Figure 11.14. If both X and Y are zero (top left), M1 and M2 are ON so they act as shorts and M3 and M4 are open, thus the output is directly connected to V_{CC}, so the output is ON.

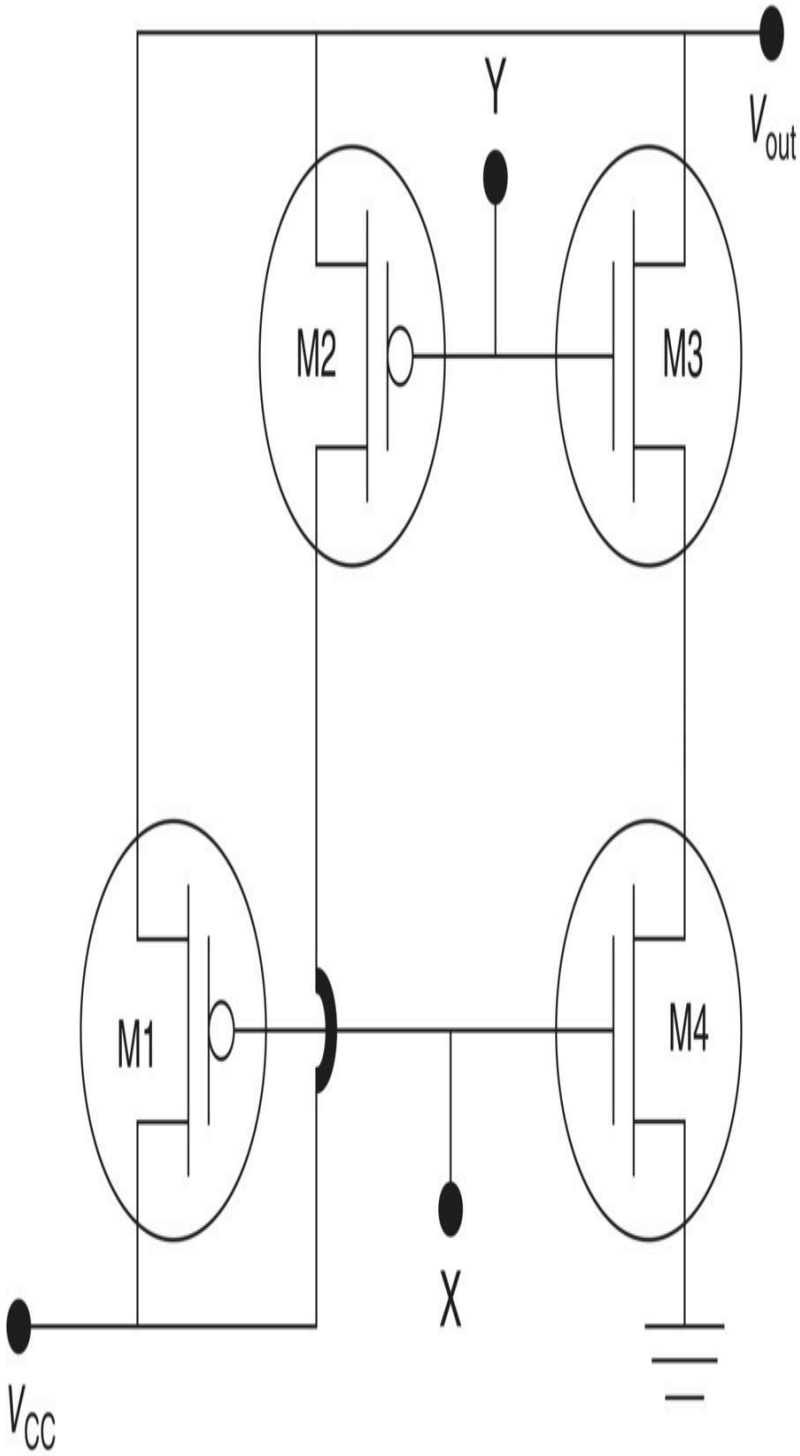
If X = 0 and Y = 1 (top right) then M1 changes from closed to open and M3 from open to closed. Now the output is disconnected from the source V_{CC} and shorted to ground through M3. The output is zero. The same happens when X = 1 and Y = 0 (bottom left) or when both X and Y are 1 (bottom right). In all of these three cases the output is shorted to ground. We have created the truth table for the NOR circuit. The output is only ON when the two inputs are OFF.

If we want to create the module OR, we just add a NOT module to the output of the NOR module and we change all the V_{out} in the truth table of Figure 11.13 from 0 to 1 and vice versa.

11.6 The NAND Circuit

As in the previous section, I discuss here the NAND circuit because it is easier to implement and understand. As before, Figure 11.15 show the CMOS implementation of the NAND circuit with the symbol and the truth table, and Figure 11.16 shows the MOSFET status as the input changes from 1 to 0.

By this time, I do not think I need to explain everything. You can follow how as X and Y change from 0 to 1, the output is connected to V_{CC} or to ground. It is only 0 when both inputs are 1. Again, if you want an AND circuit just add a NOT circuit to the NAND to get the result.



X	Y	V_{out}
0	0	V_{CC}
0	1	V_{CC}
1	0	V_{CC}
1	1	0

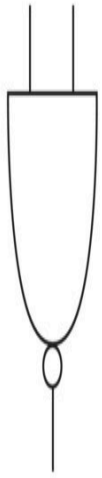


Figure 11.15 The NAND circuit (left) with the truth table (top right) and its symbol (bottom right).

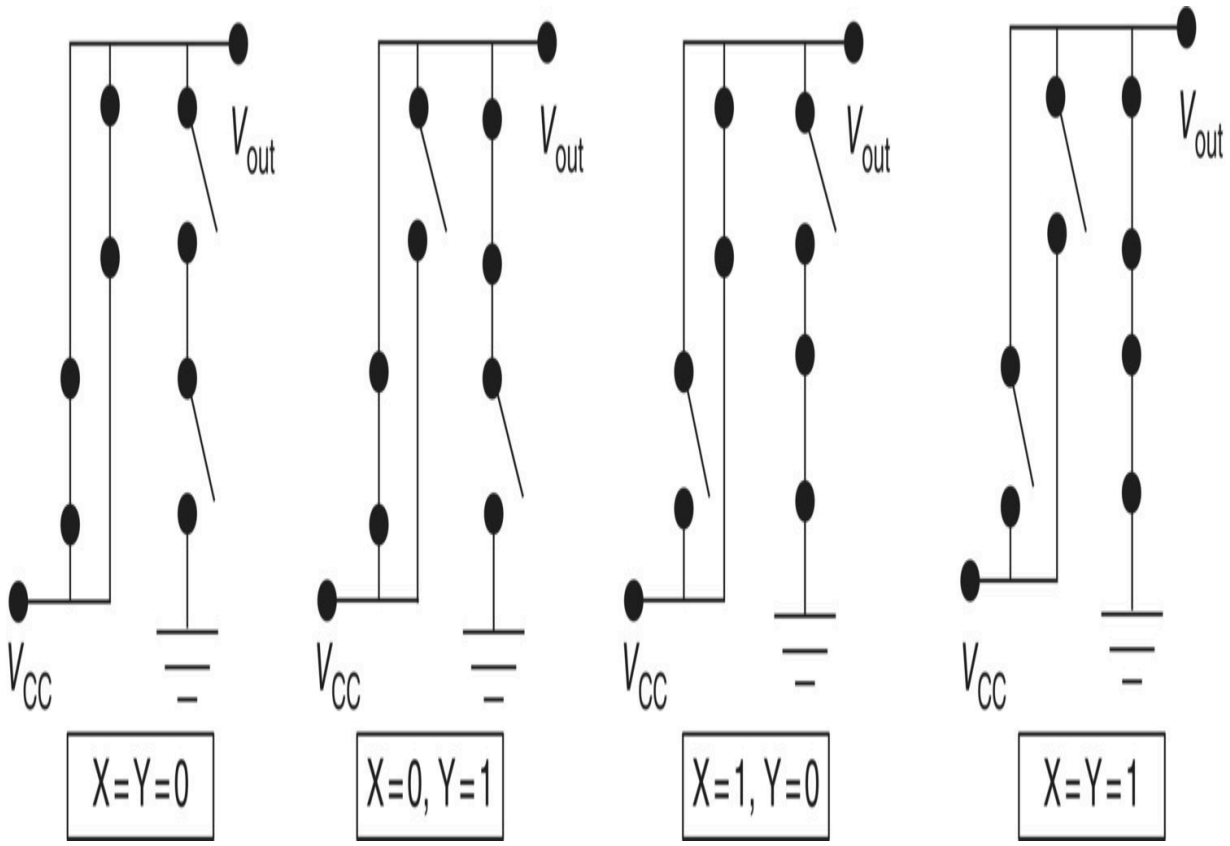


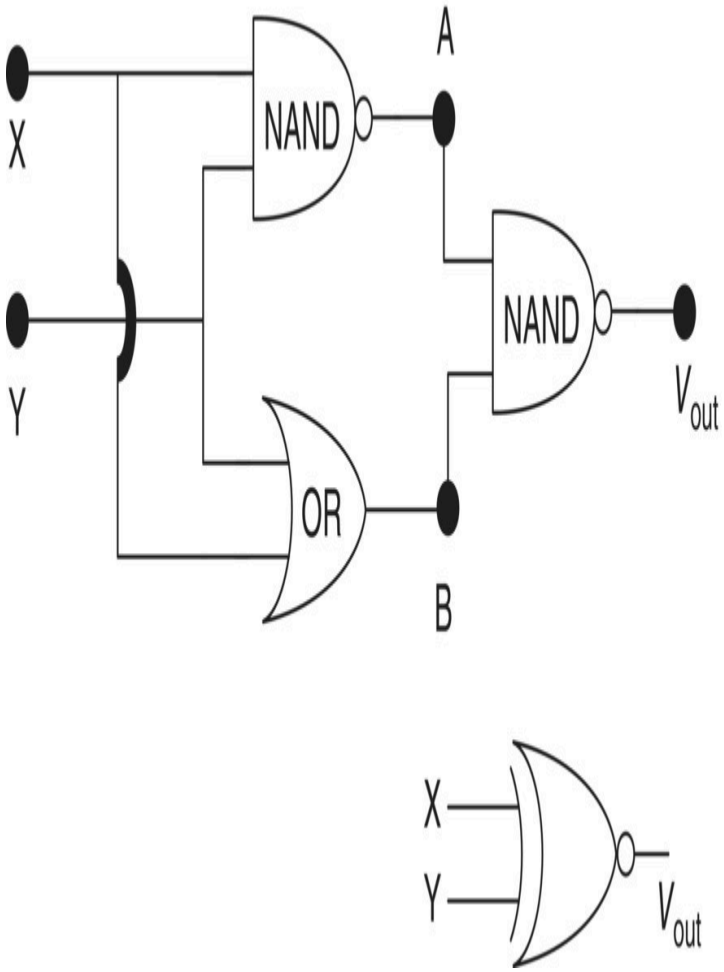
Figure 11.16 The CMOS switching status as the inputs go independently from 0 to 1. In this case only when both inputs are 1 is the output voltage grounded, 0.

11.7 The XNOR or Exclusive NOR

As I did above with the NAND and NOR circuits, I could show you another circuit consisting of CMOS that would result in the XNOR module, but now that we have found ways to represent the NAND and the OR circuits, I can construct the XNOR module by using those that we have already seen. I use the same approach later when I explain some of the arithmetic operations.

Look at Figure 11.17. For convenience I have written NAND and OR inside the logic symbols, although the symbols themselves should suffice. Also, I have added two intermediate columns in the truth table to make the explanation about how it works a little easier.

First consider point A and look only at the first three columns in the truth table. The output of a NAND is 0 only when the two inputs are 1. So, point A is 0 only when both X and Y are 1. That is what the truth table, column A states. Now look at the two input columns, X and Y, and the fourth column, the B column. The B column is the output of an OR circuit, so point B is 0 only when both X and Y are 0. Finally, points A and B are now the inputs to the second NAND, therefore, as we saw when looking at point A as an output, V_{out} is 1 only when either A or B is zero. Don't look now at columns A and B and just take a look at the first two input columns and the final output column, V_{out}. The output voltage is 1 only when either both X and Y are 1 or both are 0, exactly the negation of the XOR.



X	Y	A	B	V_{out}
0	0	1	0	1
0	1	1	1	0
1	0	1	1	0
1	1	0	1	1

Figure 11.17 The logic function XNOR, its truth table, and its symbol.

Not only we have constructed a new logic module, but you can start to see how logic circuits can be combined to get many other more complex and sophisticated operations, as we'll see in the next section.

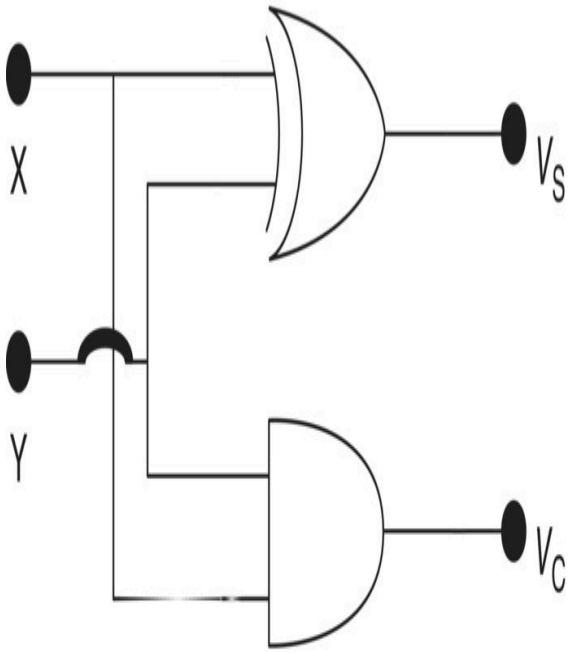
11.8 The Half Adder

In the decimal system, 4 + 8 = 12, and that is equal to 10 + 2. The 1 in front of the 2 is not a 1 but a 10. We could say that above number 9 we carry a 1 which we placed in front of the 2 to indicate that there is zero behind the 2. Similarly, in a number system based

6, 4 + 5 = 13. After 6 we carry a 1 and proceed to count so the decimal number 7 becomes 11, 8 becomes 12, and 9 becomes 13.

In the binary system there are only 1s and 0s, therefore 0 + 0 = 0, 0 + 1 = 1 and 1 + 1 = 10. We need to carry a 1 and place it in a separate column, ahead of the zero. Using the logic gates, we can now implement the addition of two one-digit operations (Figure 11.18). We call the module in Figure 11.18 a half adder because it adds only two single-digit numbers. Basically, they operate on one number at a time.

The half adder can be implemented using XOR and AND modules. If the inputs X and Y are 0, the outputs of both the XOR output, V_S for signal value and the AND output, V_C, for the carry-on result are 0. The output of the XOR, the third column in the truth table, is 1 only if one of the inputs is 1; otherwise the output is 0. The output of the AND circuit, the fourth column of the truth table, is 1 only if both X and Y are 1. This agrees with the truth table for the sum of two binary numbers, which I show in the fifth column. The symbol for a half adder is just a square box with two inputs, X and Y, and two outputs, V_S and V_C, with the word "half adder" in the middle of the box for clarity.



X	Y	V_S	V_C	Result
0	0	0	0	0
1	0	1	0	1
0	1	1	0	1
1	1	0	1	10

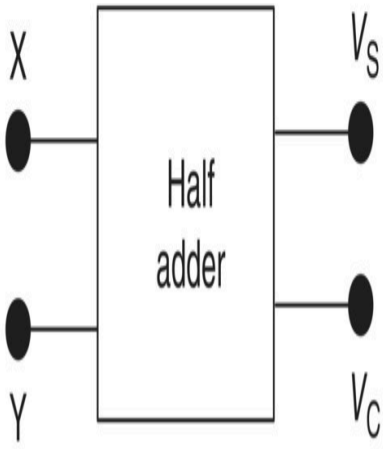


Figure 11.18 The half adder circuit (left), the truth table (right), and its symbol (lower middle).

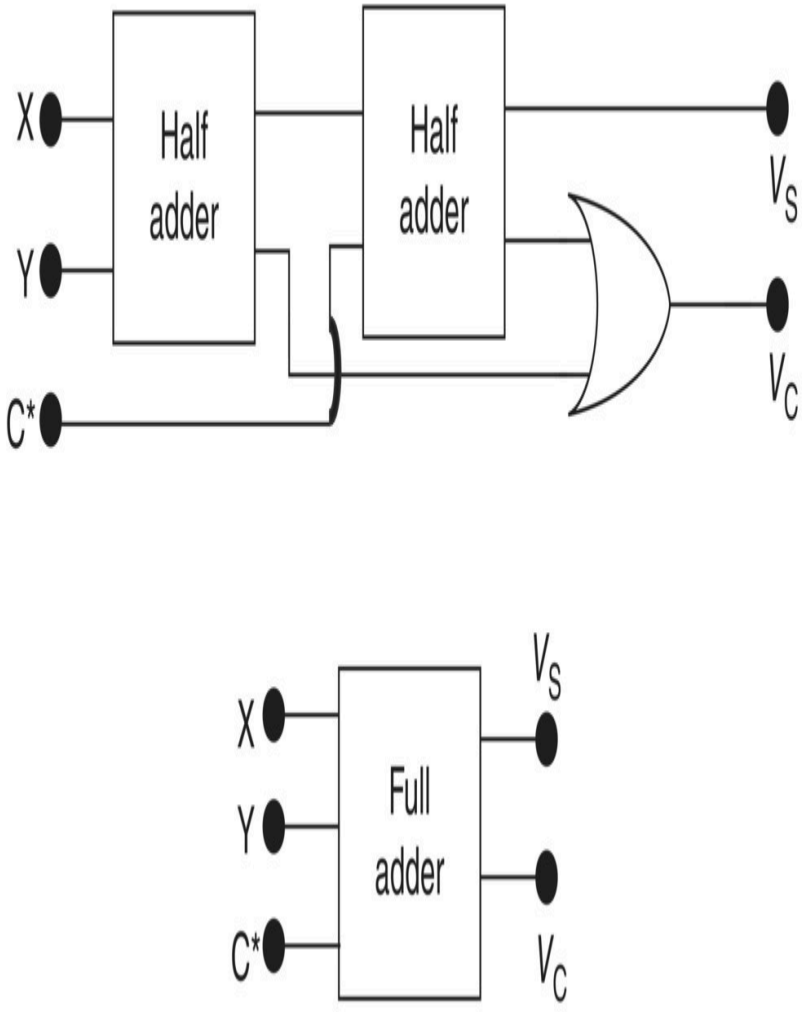
11.9 The Full Adder

The problem with the half adder is that it can add only two single-digit numbers, similar to the decimal system limited to adding just two numbers between zero and nine. What happens when we want

to add larger numbers like 15 + 3? In addition to the 5 and the 3, we have the digit 1 (which is not 1 but a 10) that we need to include in our sum. In the binary system we have a similar situation. With the half adder we can add 1 + 0 or 1 + 1 but how about 10 to 1? That is when we need the full adder, which I show in Figure 11.19.

I am using the same trick I used before. Yes, I could show you the structure of the full adder using CMOS or logic modules, but I can more easily create a full adder by using two half adders and an OR module.

Note first that the full adder has three inputs, the two digits X and Y, and the carry-on digit from a previous operation, C*. The first half adder is exactly the same as the adder I explained in Figure 11.18. The second half adder adds the carry-on of a previous operation, C*, to the output of the first half adder. Then the carry-on of the second half adder is OR with the carry-on of the first half adder. So, the new carry-on V_C is 1 when two or all three inputs are 1 and V_S is 1 when just one or all three inputs are 1. See if you can follow the logic. I explain the logic, step by step, in detail in Appendix 11.2.



Three inputs			Two outputs	
X	Y	C*	V_S	V_C
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Figure 11.19 The full adder with the truth table and the new symbol can be constructed from two half adders and an OR circuit.

11.10 Adding More than Two Digital Numbers

In the great majority of arithmetic operations, we want to add more than just two one-digit numbers. Figure 11.20 shows how to add two three-digit numbers. The resulting sum may have not three but four digits, depending on the last carry-over. That is what the carry-on, C4, does. If you have more than three digits or more than two numbers, you add more and more full adders to the chain. (Note

that I reversed the direction of the sum in Figure 11.20 from right to left, because it is much easier to draw it.)

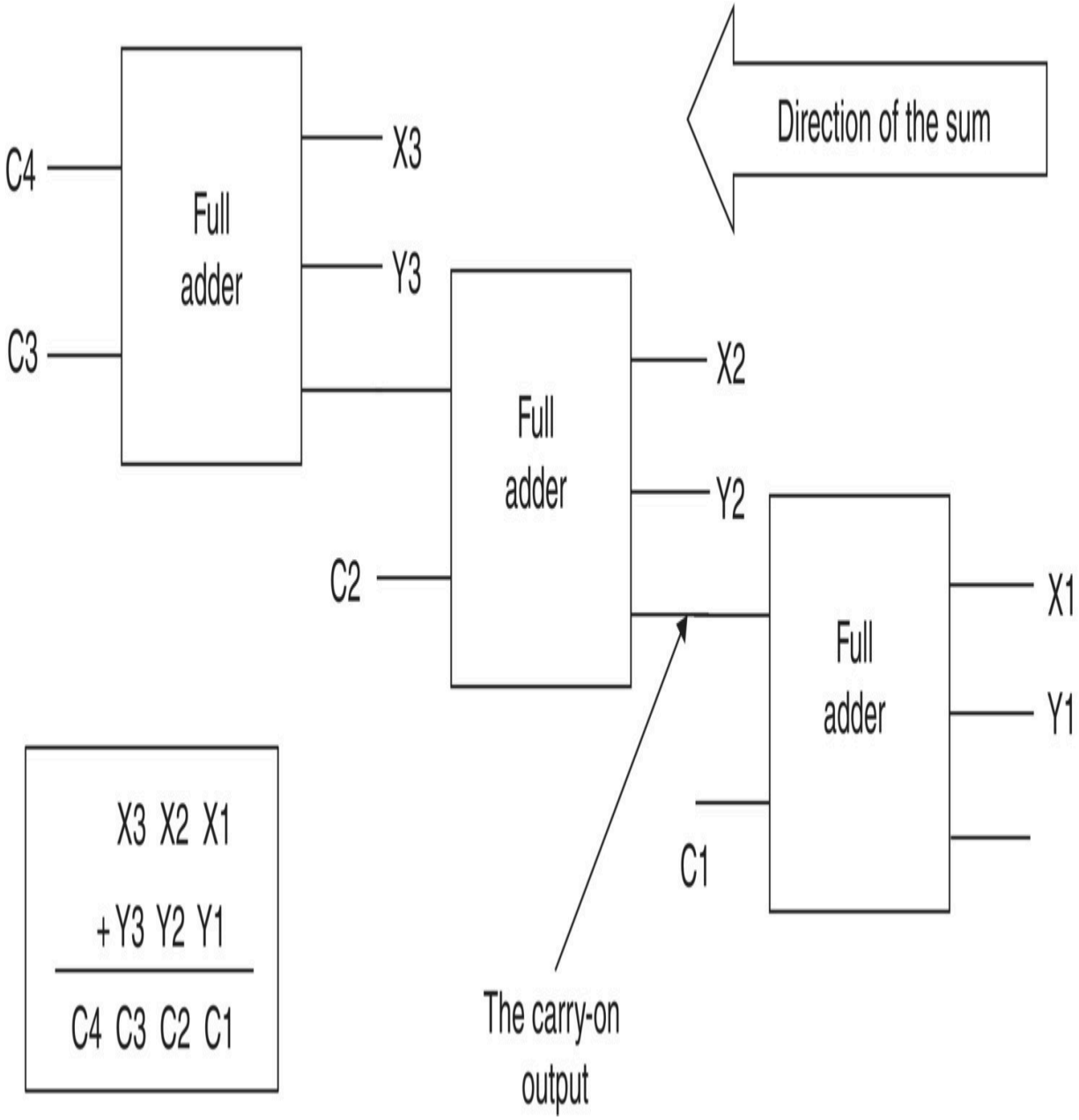


Figure 11.20 Adding two three-digit numbers. We use as many full adders as the number of digit inputs we want to add.

11.11 The Subtractor

To see how a computer subtracts two numbers we have to go back to the way children learn how to subtract. The child in elementary school tries to subtract 7 from 5 (Figure 11.21A) but he can only do it if he changes the number 5 to a 15 (Figure 11.21B). He gets an 8. Then he increases the number 3 in the subtrahend by 1 to 4 (Figure 11.21C) and subtracts it from 3, which again he can do as long as he makes it 13 (Figure 11.21D). He writes a 9 and changes the 1 in the subtrahend to 2 (Figure 11.21E) and he gets the final result, a 0 (Figure 11.21F). The high school student can do this same operation in his head (the substraction on the right).

Let's now subtract two digital numbers (Figure 11.22). Suppose we want to subtract 38 from 59 in digital numbers. The operation is 111011 minus 100110. Let me do this slowly. In block A, I subtract the last two numbers. These are easy, 1–0 = 1 and 1–1 = 0. We have a problem with the next digits. Now we have to subtract 1 from 0, which we cannot do. So, we borrow the 1 from the left digit of the minuend so the third digit from the right becomes 10 and the fourth digit from the right loses the 1 and changes to 0. I show this in block B. Digital 10 is number 2 in decimal, so in digital 10 – 1 = 1. The result of the subtraction of the third digits is 1. The fourth digits are now 0 in the minuend and in the subtrahend, so the result is 0. The remaining two digits on the left are 1 – 0 = 1 and 1 – 1 = 0 (block C), completing the subtraction (block D).

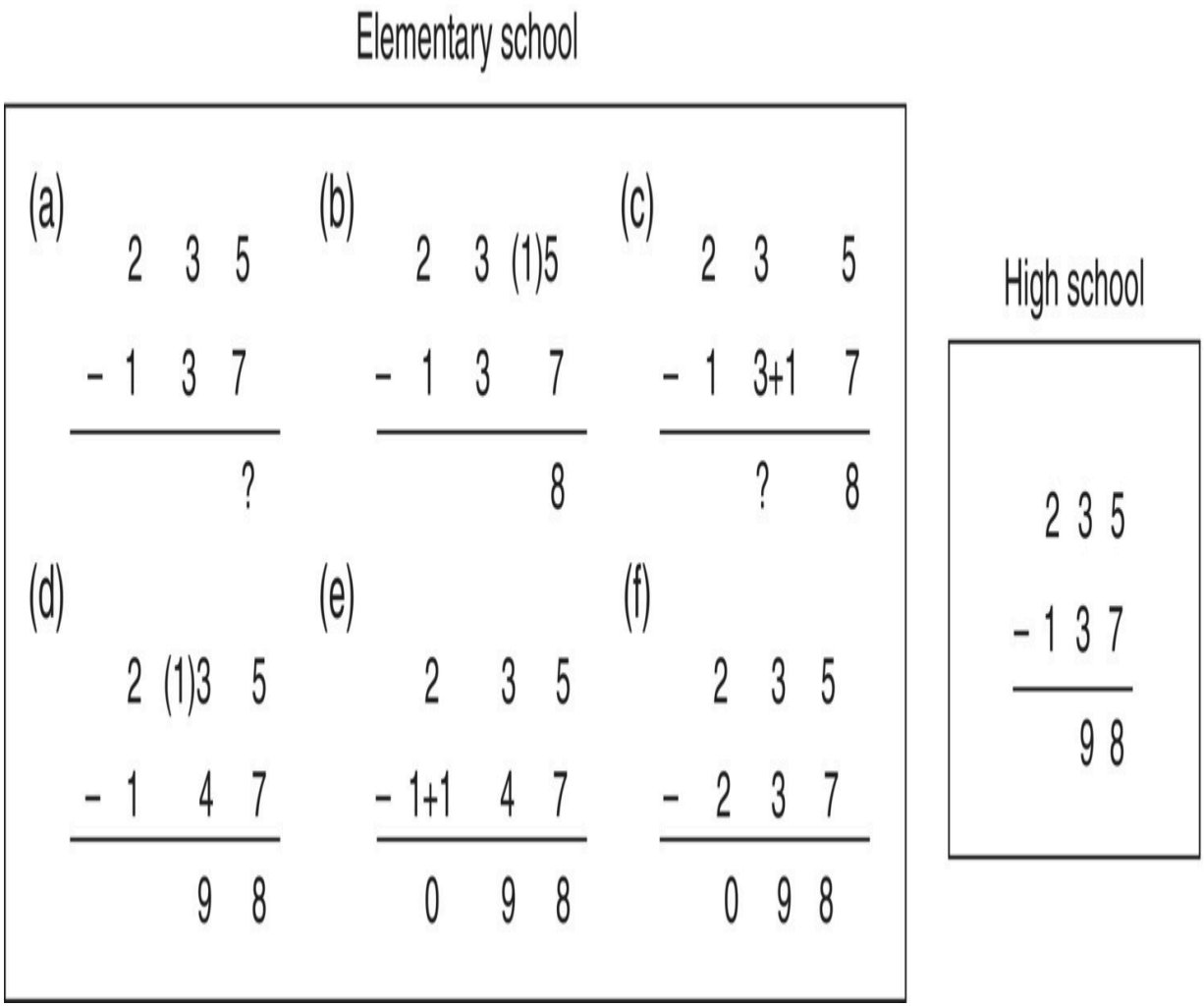


Figure 11.21 How elementary (left) and high school students (right) subtract two numbers.

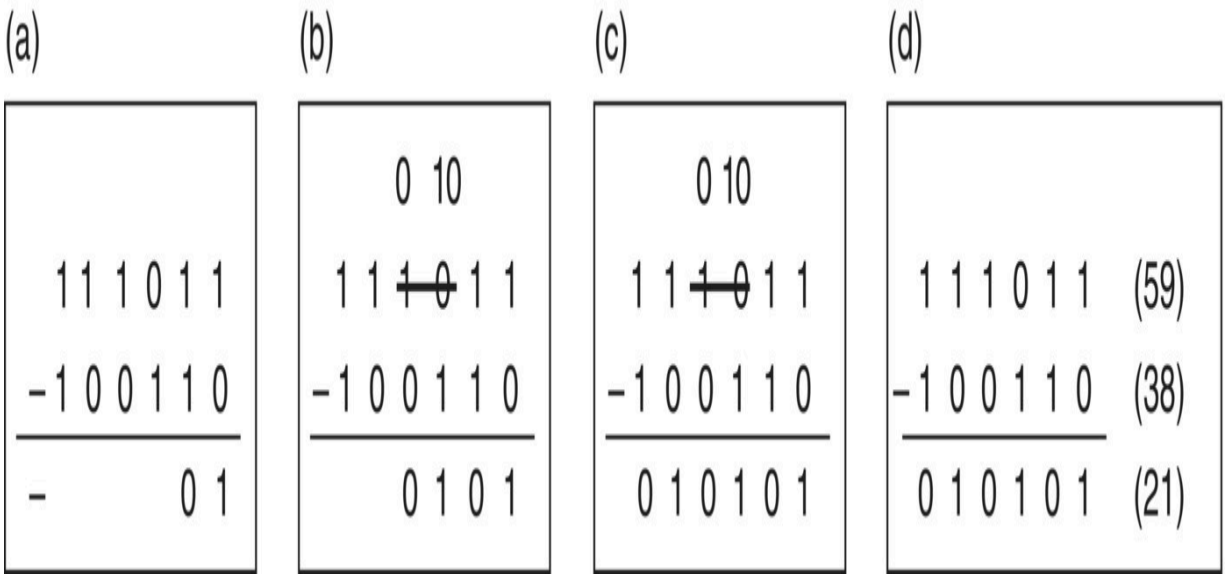
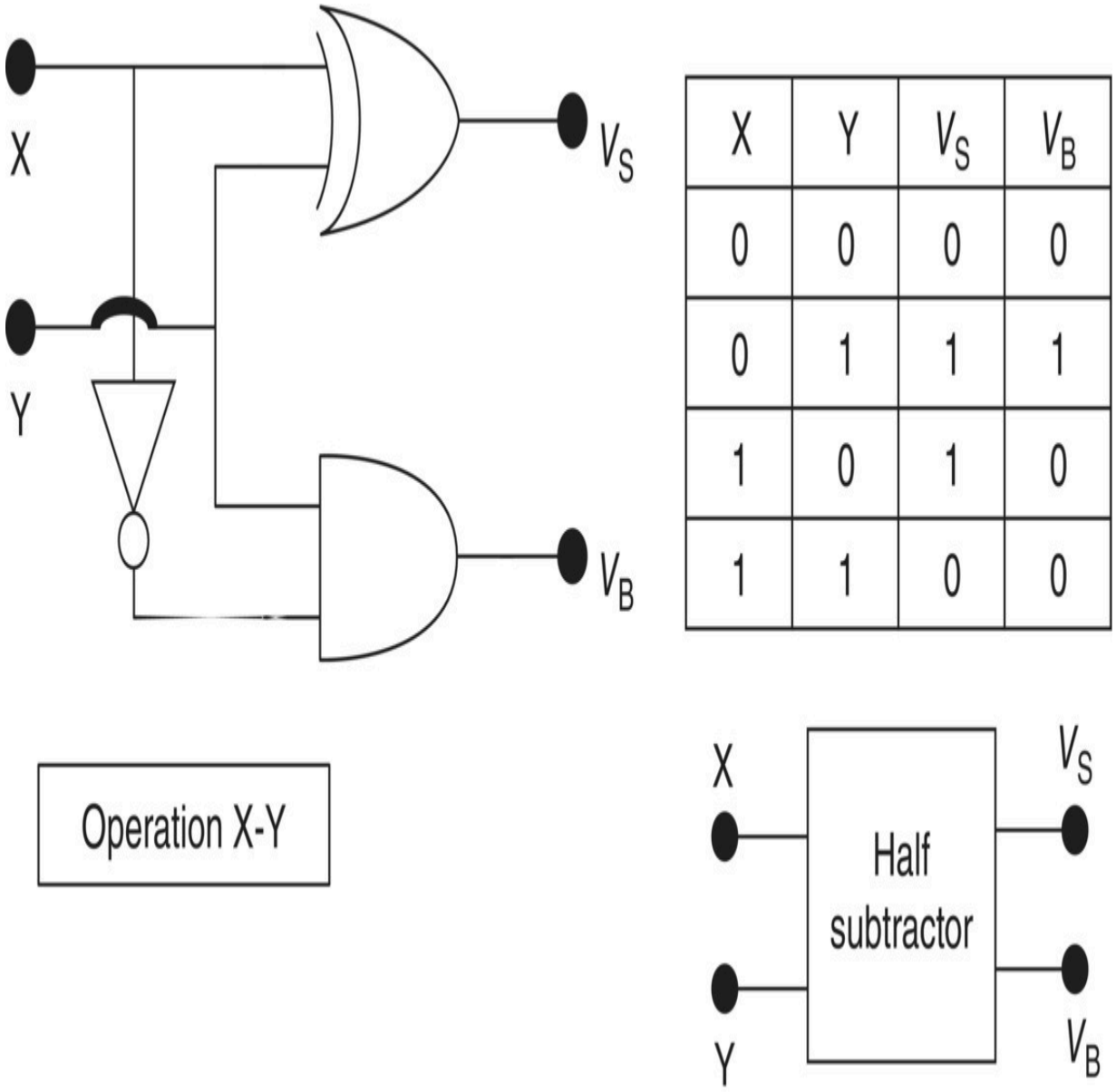


Figure 11.22 Step-by-step subtraction of two digital numbers.



X	Y	V_S	V_B
0	0	0	0
0	1	1	1
1	0	1	0
1	1	0	0

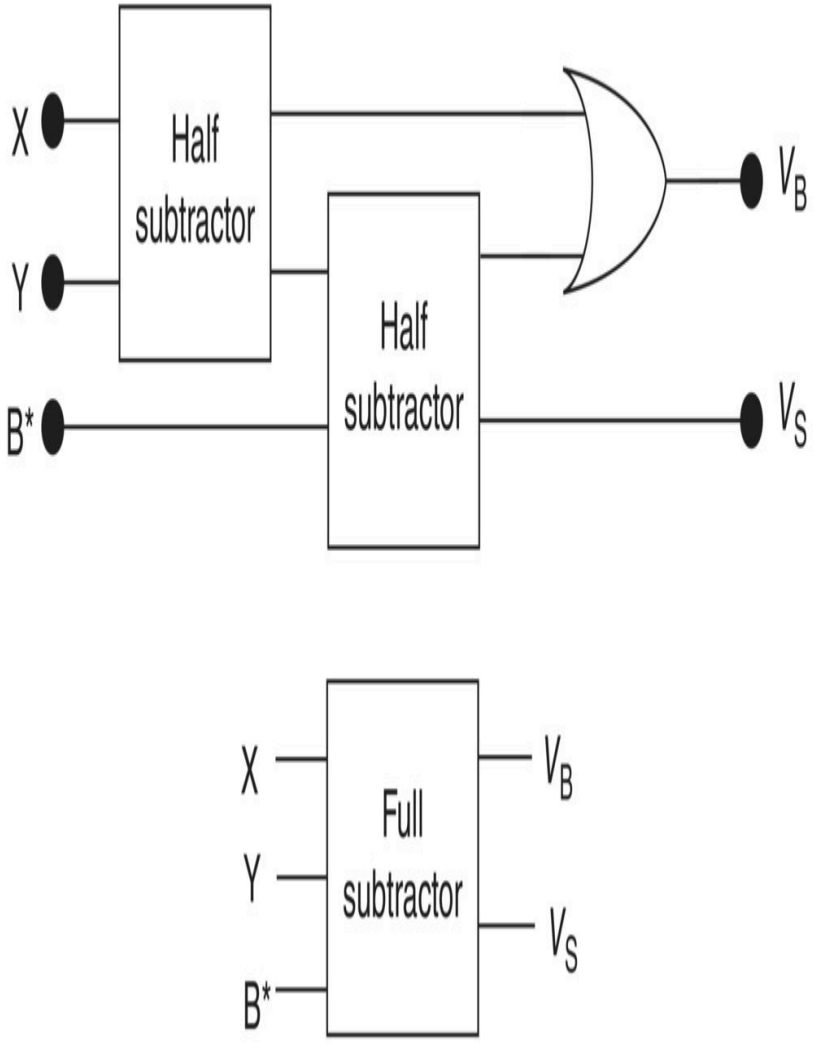
Figure 11.23 The half subtractor circuit (left), the truth table (top right), and its symbol (lower right).

Another way of performing digital subtractions is to use complementary numbers, and this is actually easier to implement in a digital circuit. I cover subtraction using complementary numbers in Appendix 11.3.

A computer is not as smart as a high school student and does the substation with binary numbers in the same way as the child in elementary school. Figure 11.23 shows the circuit implementation of

the half subtractor and the truth table. It is almost identical to the half adder in Figure 11.18. The only change is that I have added an inverter, a NOT module, between the input X and the AND module. The output signal, V_S, is the same as the adder, that is, it is 1 when either X or Y are 1 but not both. That is exactly what the XOR module does. The output of the AND module is the "borrow" digit, V_B. which is only 1 if the subtrahend, Y, is larger than the minuend, X. The NOT module changes the value of X so the only time that the AND circuit sees two 1s is when X = 0 and Y = 1, which precisely what we want. This is what I show in the truth table for the operation X – Y.

If you want a full subtractor you'll do the same thing we did with the full adder, with minor changes (Figure 11.24).



Three inputs			Two inputs	
X	Y	B*	V_S	V_S
0	0	0	0	0
0	0	1	1	1
0	1	0	1	1
0	1	1	1	0
1	0	0	0	1
1	0	1	0	0
1	1	0	0	0
1	1	1	1	1

Figure 11.24 Full subtractor (top left), its symbol (lower left), and the truth table (right).

The OR circuit that in the adder gave the result for V_C now gives the result for V_S, and vice versa, and that is all it changes. In Appendix 11.3 I discuss complementary numbers which help us to understand why we subtract two digital number by adding one to the complement of the other.

11.12 Digression: Flip-flops, Latches, and Shifters

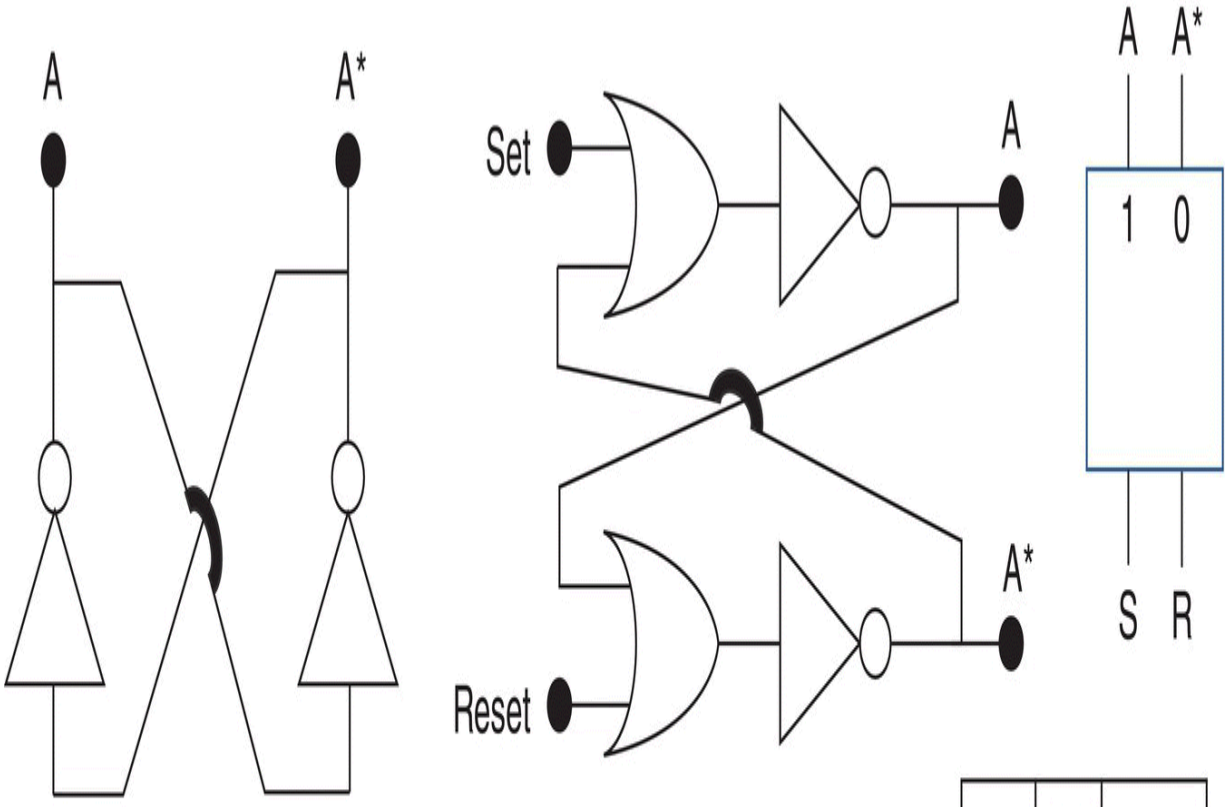
To explain other arithmetic operations, like multiplication of digital numbers, we need to understand a way of shifting numbers. To do the shifting we need two different operationally similar circuits, the flip-flop and the latch. Figure 11.25 shows a flip-flop on the left and the latch on the right.

These circuits look strangely similar to two dogs biting each other's tails. Take a look first at the figure on the left. It consists of two NOT modules. If A is 1 then A* must be 0 and vice versa, if A is 0 then A* is 1, which is very stable in both cases. We can use this configuration to store a digit. Although it is arbitrary, we select one input, in this case A, to define the state of the flip-flop. We say the flip-flop is 0 when A is 0 and, vice versa, 1 when A is 1. I show this in the truth table at the lower right of Figure 11.25. If I add two OR modules, one at each input of the NOT circuits, then I have a way of changing A and A* from 0 to 1 or from 1 to 0. Once I have done that, A and A* remain in this status, unchanged, until I decide to change them by using the reset input. We call the circuit in the middle a latch. Now we have a circuit, a latch, that allow us to keep a value, 1 or 0, at output A, for as long as we want. I show the symbol for a latch on the right of Figure 11.25. The 1 and the 0 can be interchanged depending on the status of the latch.

The next circuit I want to discuss, because we need it to understand multiplication, is the shift operation, also known as the rotation operation. Figure 11.26 shows a three-input shift circuit (Note that for convenience I show connecting points with black dots so if two lines cross each other without a dot, it means that they are not connected.)

The circuit has inputs on the left, outputs on the right, and the registers on the top, in our case just three, but there can be hundreds. Each register is connected to all the gates of a vertical column. The three inputs are connected to the source of the CMOS

in each horizontal row. The connection of the collector of each of the CMOS to the output line is a little trickier. Look at the first column, R1, of CMOS. The input I1 is connected through its first CMOS to output O1. Similarly, I2 and I3 are connected to O2 and O3 through the respective CMOS of the first column.



A	A*	State
0	1	0
1	0	1

Figure 11.25 Flip-flop (left) and latch (middle) modules, their symbol, and the truth table (right).

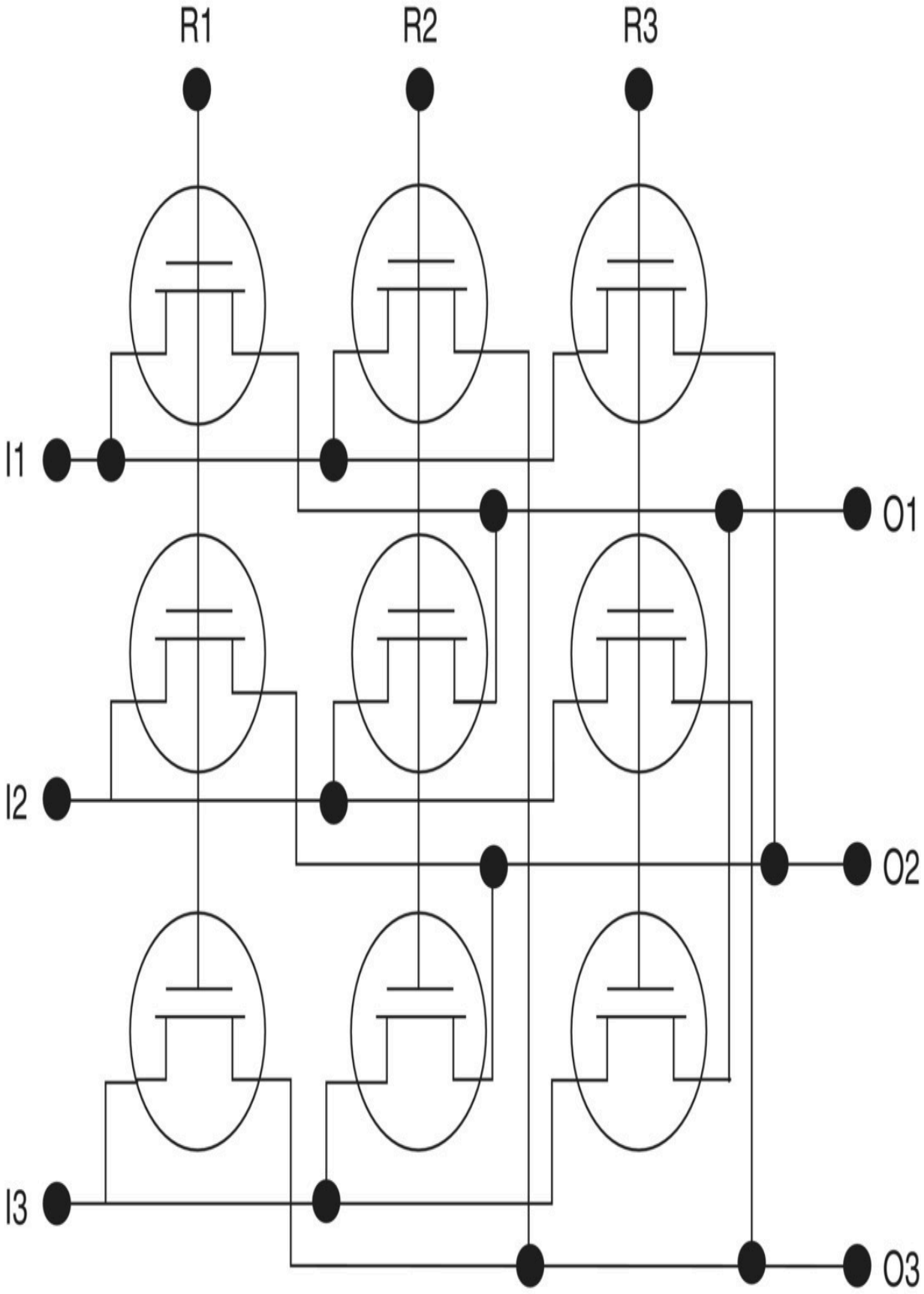

R1
R2
R3
I1
I2
I3
O1
O2
O3

Figure 11.26 A 3 × 3 shift register.

The connections to the collector of the CMOS of the second column, R2, are shifted; the collector of the top CMOS on the second column is connected to O3, the second to O1, and the third to O2. Finally, let's look at the collectors of the third column, R3. The first CMOS is connected to O2 the second to O3 and the third of O1. I should mention that if we had a 10 × 10 matrix, the first MOS of the second column would be connected to the tenth, the last, output, the second to the first, the third to the second, etc. The first transistor of the third column would be connected to the ninth output, the second to the tenth, the third to the first, etc. I think you can see the pattern.

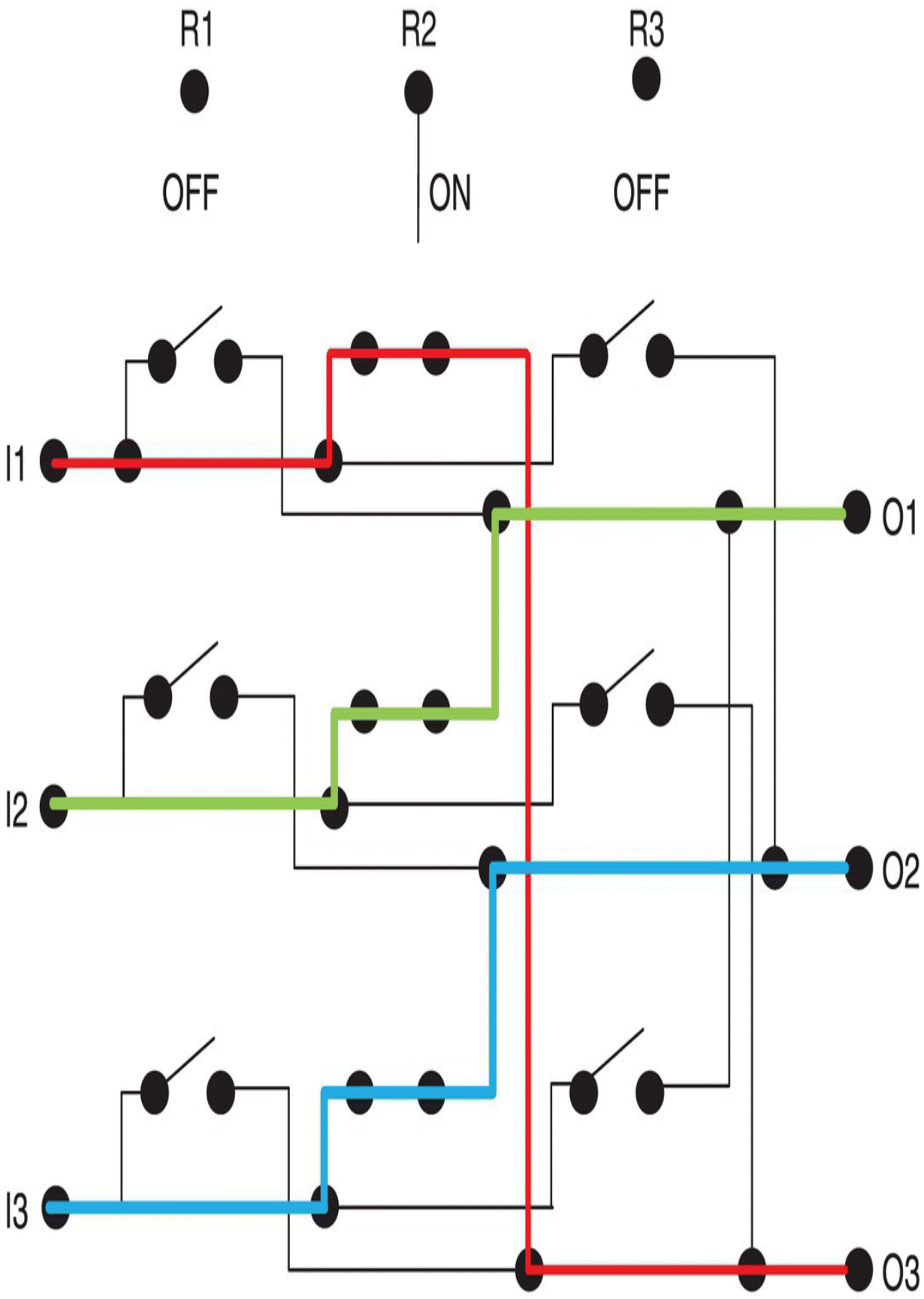

R1
R2
R3
OFF
ON
OFF
I1
I2
I3
O1
O2
O3

Figure 11.27 Electrical path of Figure 11.26 when R2 is ON and all the others are OFF.

Now go back to our 3 × 3 example. Remember that we use these CMOS as switches. Assume, for example, that we turn ON R2. Using switches, the circuit of Figure 11.26 looks like the one in Figure 11.27.

All the CMOSs of the second column are ON and all the others are OFF. If you follow the path of all the closed switches (in bold lines), you'll see that the first output is connected to the second input O1 = I2, the second output to the third input, O2 = I3, and the third output to the first input, O3 = I1. We have shifted all the outputs one step behind the input. You can also see why we call this circuit also a rotation circuit.

Now we are ready to discuss the multiplication operation.

11.13 Multiplication and Division of Binary Numbers

Take a look at the truth table of the AND module in Figure 11.3. The output of the AND circuit is 1 only when both binary digits are 1, ON, exactly what multiplication does. Any digit multiplied by zero is zero and the only time the result will be 1 is when both digits are 1 (1 × 1 = 1), therefore just the simple AND circuit multiplies two digital numbers. What is more involved is when we want to multiply two numbers with more than 1 digit each. Let us say we want to multiply 110101 (equivalent to 53 in the decimal system) by 1101 (equivalent to 13) (Figure 11.28). What we do manually is multiply the upper number, the multiplicand, by the first digit of the second number, the multiplayer, and then we multiply the second number of the multiplier with the multiplicand, shift this second product, and add the two results (left of Figure 11.28). We do the same with the digital multiplication, shown in the middle block. Since the first digit of the multiplier is 1, the first result is just the same as the multiplicand. Then we take the second digit of the multiplier, which

in our case happens to be zero and therefore the results of multiplying this second digit, zero, by the multiplicand is all zeros. But now we have shifted the result one position to the left. The third digit of the multiplier is also 1 so we just write the multiplicand in the third row, shifting the numbers one more position to the left, etc. Finally we add all the partial multiplication results.

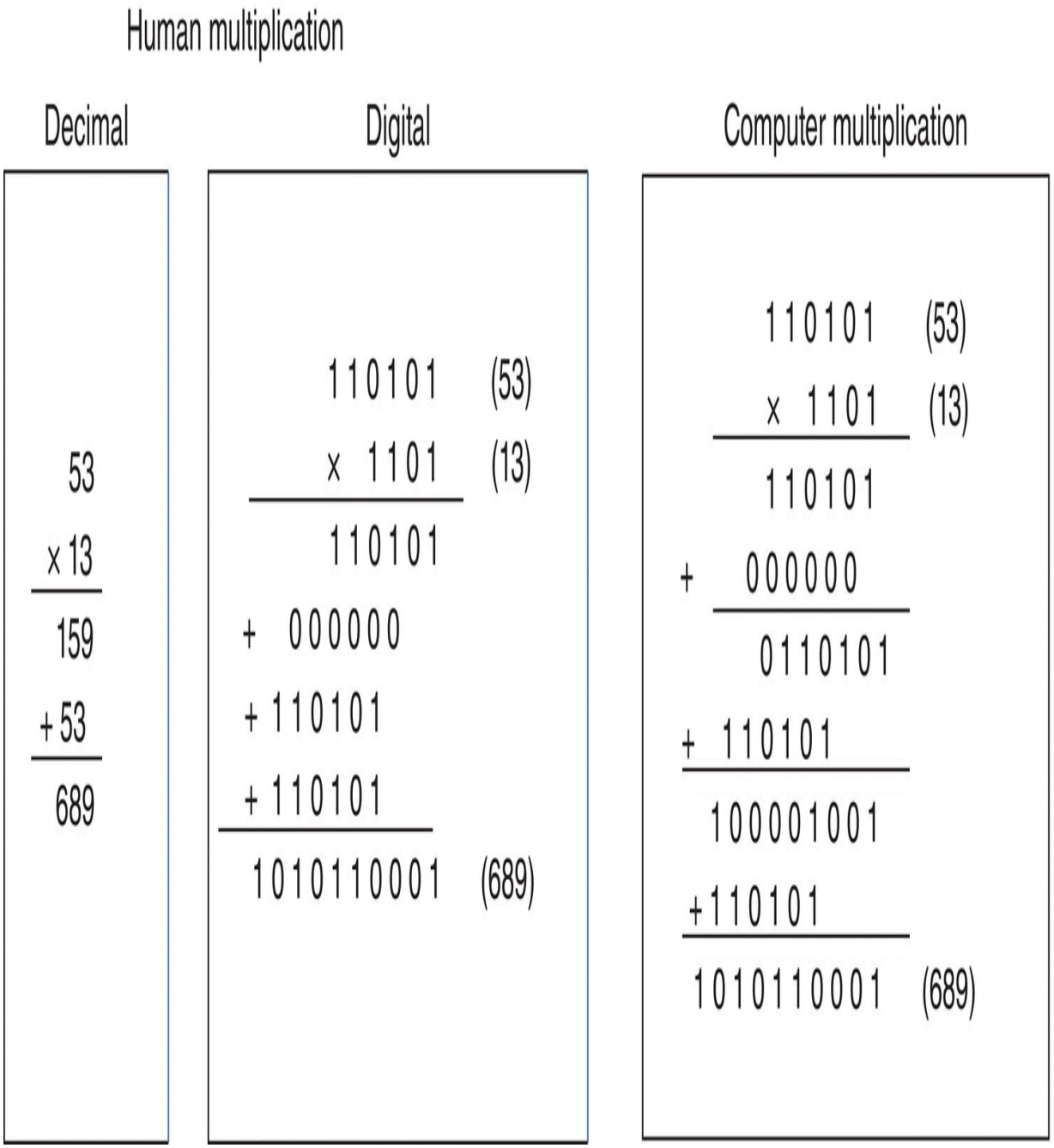


Figure 11.28 The multiplication of two digital numbers is the same as in the decimal system, multiplying each digit of the multiplicand to the multiplier and shifting the product one. The computer can only add two numbers at a time, so it requires additional steps.

A computer has difficulty adding more than two numbers at the same time, so it does this by adding the first two numbers and then this result is added to the third number and then the new result to

the fourth number, so the computer operation looks like the box on the right in Figure 11.28. The computer stores the results of each partial sum in an accumulator, a bunch of latches, then the next product is shifted and added to the previous result and sent back to the accumulator etc. until it finishes multiplying (or adding and shifting) all the digits.

In the same way that we perform multiplications by adding and shifting each intermediate result, we divide two numbers by subtracting and shifting the result, and then keep subtracting and shifting until all the digits are accounted for (see Appendix 11.4).

11.14 Additional Comments: Speed and Power

As you might expect there are many other ways to obtain logic modules from CMOS and junction transistors than the ones I have shown you here. Many of these other combinations may be desirable because of the speed or the power or even the circuit layout configuration that we need. You can find all of these options in many technical books, but I hope that this chapter has given you an idea of how engineers use semiconductor devices to create complex arithmetic and processing systems.

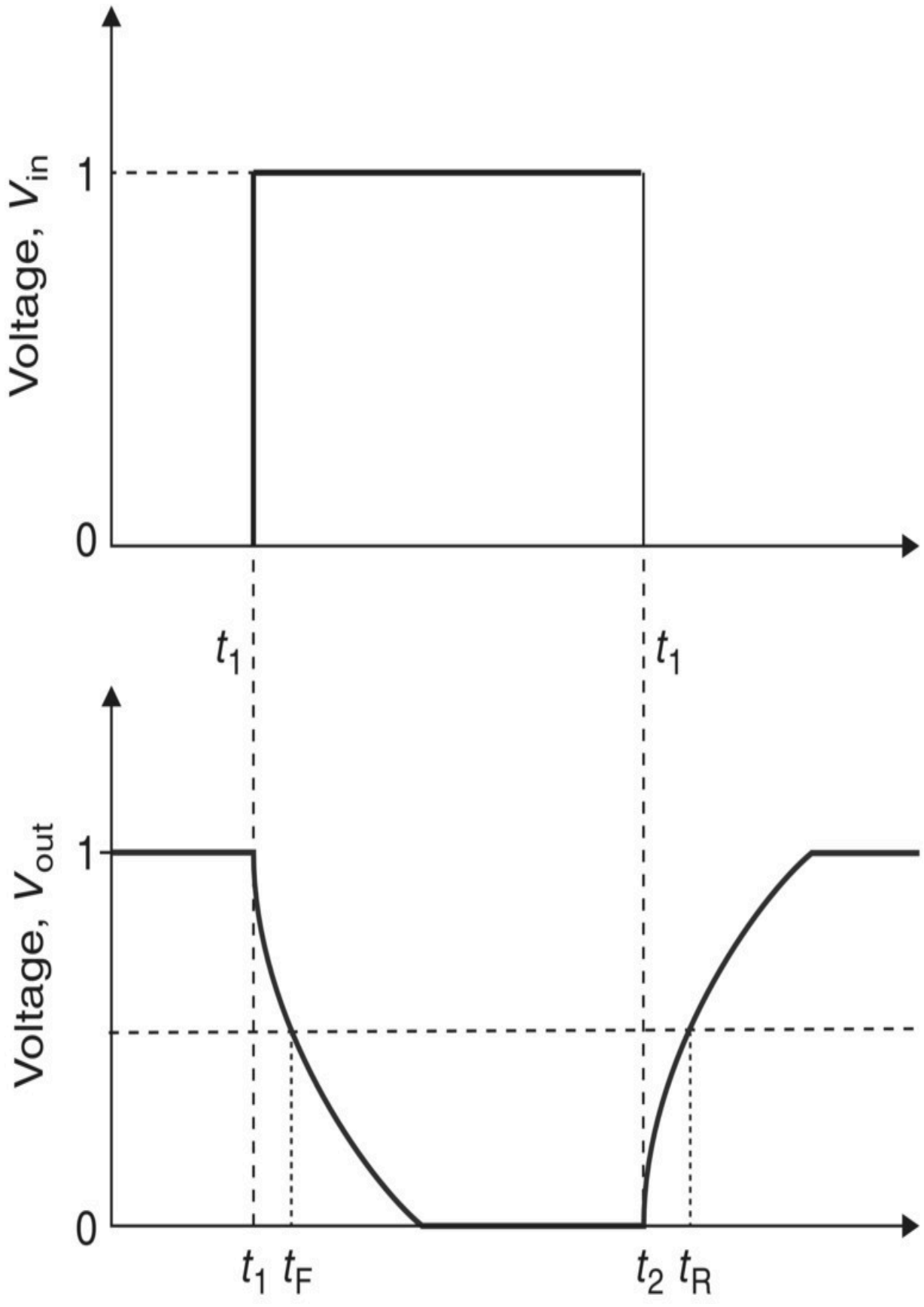

Voltage, V_{in}
1
0
t_1
t_1
Voltage, V_{out}
1
0
t_1 t_F
t_2 t_R

Figure 11.29 The output of a device driven by a perfect square input pulse (top) will have a rise and a fall time and will not have the instantaneous rise or fall of the original input (below).

Another comment I would like to make is about the speed of these circuits. We want fast logic to perform complex operations in the shortest amount of time. This is usually one of the most important layout and circuit designs considerations that the engineer has to deal with. Take a look at Figure 11.29. Suppose that the input voltage, V_{in}, in a circuit such as a NOT module, for example, changes instantaneously from 0 to 1, as I show in the upper part of the figure. We would like the response time to be so fast that the output voltage, V_{out}, goes from 1 to 0 instantaneously. The reality is that the output voltage takes some time before the CMOSs or transistors change from 1 to 0 (lower part of the figure). The time it takes for the output to go from 1 to 0 is the fall time. The rise time at the other end, when the input goes from 1 to 0, is $t_R - t_2$, and we call this the rise time. There are lots of capacitive and resistive effects due to the lines and the properties of the semiconductor devices themselves. These effects slow down and limit the speed of the operation. We can define a propagation delay, t_R and t_F, by considering when the signal reaches 50% of its final value (any decaying function takes forever to reach its final value, so the 50% measure makes sense to get an idea of how fast our systems are). In any electronic circuit there are thousands of transistors, so propagation delays are one of the most critical design considerations.

Another point I want to discuss is power dissipation. In digital switching circuits, we consider two sources of power: DC power and switching power.

Let's consider first the DC case. When CMOS devices are OFF, there is no DC current, except for a tiny leakage current. When they are ON, the power is the source voltage times the current through the device. Nevertheless, most switching systems consist of both n- and

p-type MOS in the circuit, as you can see in Figures 11.11–11.16, so when one of them is ON the other is OFF. So, most DC currents are basically due to leakage currents.

The switching power is due to all the capacitances in the circuits. Some capacitances are just the parasitic capacitances due to lines crossing other lines and the semiconductor devices themselves. When I change the status of a circuit, turn it ON for example, there is a fast movement of charges to the capacitances. When I turn it OFF, the charges stored in the capacitances have to discharge. The power dissipation is proportional to the charge stored and the frequency of the switching. The circuit layout engineer has a big job trying to minimize these parasitic capacitances.

11.15 Summary and Conclusions

In this chapter we have looked at the operations that circuits fabricated with semiconductor components allow us to perform. All are based in Boolean algebra and symbolic symbols to represent the key functional operations OR, AND, and NOT, and variations and negations of them. We have seen that these Boolean functions allow us to do digital arithmetic calculations, additions, subtractions, multiplications, and divisions, using ones and zeros. Everything that the computer does, from writing a letter to calculating the trajectory of a planet or next week's weather, is based on these simple devices. Now on to more complex devices.

Appendix 11.1 Algebraic Formulation of Logic Modules

Engineers like to express the truth function using equations rather than tables. This helps them to select circuits or modules that perform these functions. Take, for example, the sum operation. If I use a bar on top of a letter to indicate the negative of the letter, then I can express the formula for the sum of two digital numbers by

$$\text{sum} = \bar{A}B + A\bar{B}$$

$$\text{carry-on} = AB$$

If A is 1 and B is 0, or vice versa, one term of the sum is 1. The first product nonA times B is $0 \times 0 = 0$ and the second product A + nonB is $1 \times 1 = 1$, so the sum is $0 + 1 = 1$. The carry-on is zero. If both A and B are 1 then the sum is 0 and the carry-on is 1. If A and B are 0, both the sum and the carry-on are 0, which represent algebraically the same conditions as the truth table.

The formula for the difference is

$$\text{difference} = \bar{A}B + A\bar{B}$$

$$\text{borrow} = \bar{A}B$$

These equations help the designer select the combination of modules he wants to use. This formulation is also used by programmers to write their algorithms.

You can imagine that as circuits get more and more complicated the designer uses computer tools to create them. One of the most commonly used programs is Verilog, which is a hardware description language (HDL). It works around components, the same way as we did, by creating larger modules. We do not need to worry about what is inside each module. The components are defined by their inputs and outputs.

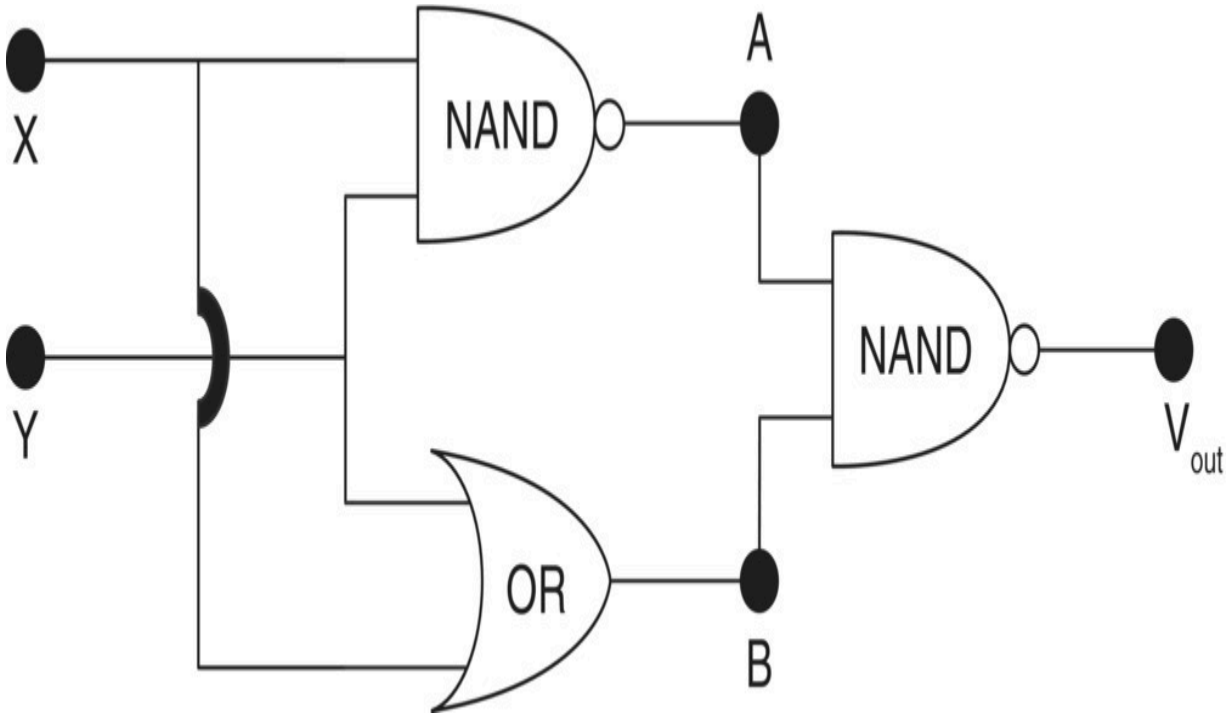


Figure 11.30 The half adder module.

Suppose, for example, we want to design the module for the half adder I show in Figure 11.18, which I repeat it here for convenience (Figure 11.30). I call this module "halfadder." It consists of one output, two inputs, and two NANDs and one OR modules. (We call these submodules or "primitives" because they address the original, primitive, Boolean algebra functions that I discussed in the first sections of this chapter.) Now it is just a question of identifying and connecting theses inputs and outputs. The computer program will look something like this:

module halfadder (Vout, X, Y)	Names the module that we want to create
input X, Y	Tells the computer what your inputs are
output Vout	Tells the computer what your outputs are
wire A, B	Identifies the internal points
nand M1 (A, X, Y)	The first NAND module has an output A and inputs X and Y

or M2 (B, X, Y)	The OR module has an output B and inputs X and Y
and M3 (Vout, B, A)	The second NAND module has an output V_{out} and the inputs A and B
endmodule	and the module is complete

You can see that the programing is rather intuitive. This same computer software can simulate and predict the operation of the half adder you have just designed. When the module is validated, you can use it any time you need it, with no need to reinvent the wheel or the module.

There are many more modules already built into the program that you can select to make it easier for you to do the design. Also, the program calculates the rise and fall times and the power dissipation. When you have finished and have validated the entire design, you can send the program to the manufacturer for him to use to design the processing masks and calculate process times and temperatures.

Appendix 11.2 Detailed Analysis of the Full Adder

Figure 11.31 shows step by step the truth table of the full adder. I like to believe that you have already figured it out. I have added three points to the figure, points A, B and C. Table 1 shows the truth table of the first half adder. As we saw in Figure 11.18, the signal A is 1 if and only if one of the inputs is 1 and is 0 otherwise. That is what column A is telling us. The carry-on, Colum B, is 1 only when both inputs are 1.

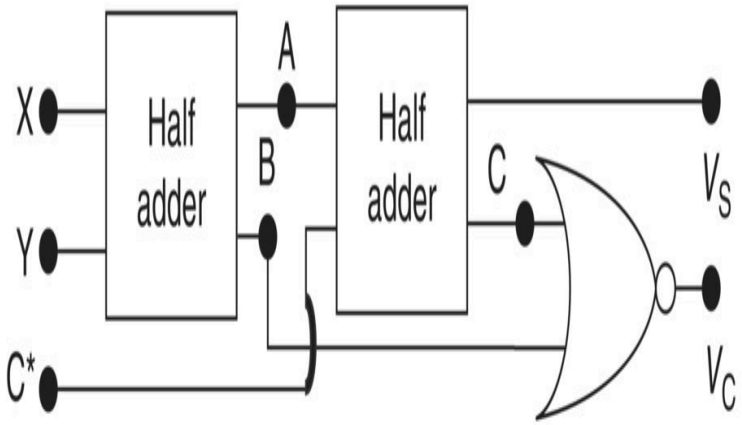


Table 1

X	Y	A	B
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

Table 2

A	C*	V_S	C
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

Table 3

C	B	V_C
0	0	0
0	1	1
1	0	1
1	1	1

X	Y	C*	A	B	C	V_S	V_C
0	0	0	0	0	0	0	0
0	0	1	0	0	0	1	0
0	1	0	1	0	0	1	0
0	1	1	1	0	1	0	1
1	0	0	1	0	0	1	0
1	0	1	1	0	1	0	1
1	1	0	0	1	0	0	1
1	1	1	0	1	0	1	1

Figure 11.31 The development of the truth table of the full adder.

Now we look at the outputs of the second half adder in Table 2. It is exactly the same, except that now we add the signal of point A to the carry-on signal, C*, from a previous operation.

Table 3 is the truth table of an OR function with the inputs B and C from Tables 1 and 2. The output V_C is 1 only when at least one of the inputs is 1, otherwise it is zero. If now you combine the three partial truth tables, we get the truth table of a full adder.

Appendix 11.3 Complementary Numbers

If you add one digital number to the complementary of another digital number you actually obtain the difference of the two numbers. Let me clarify this. In digital numbers the first bit is used to determine if the number is positive or negative. If the first digit of a number is a 0, the number is positive. If it is 1, it is negative.

Adding a number to the complementary of another number is the same as subtracting the second number from the first one. What is a complementary number? It is the negative of the number plus one. Let me show you. Suppose you want to subtract 1101 (number 13 in the decimal system) from 10100 (number 20 in the decimal system). First, we take the complementary of 13, that is we take 1101 and change all the 1s into 0s and vice versa, and then we add 1. So, 1101 changes first to 0010 and when we add 1, we get 0011. Now if we add 20, that is, 10 100, to the complement of 14, 0011, we get 10 111 and ignoring the first digit we have 0111, which happens to be 7. Let's try another subtraction, 118 – 46 = 72.

118 in digital is	1110110
and 46 is	101110
Complement of 46 is	010001
add 1	010010
add 118	1110110
to the complement of 46	+<u>010010</u>
The result is	1001000
which happens to be equal to 72	

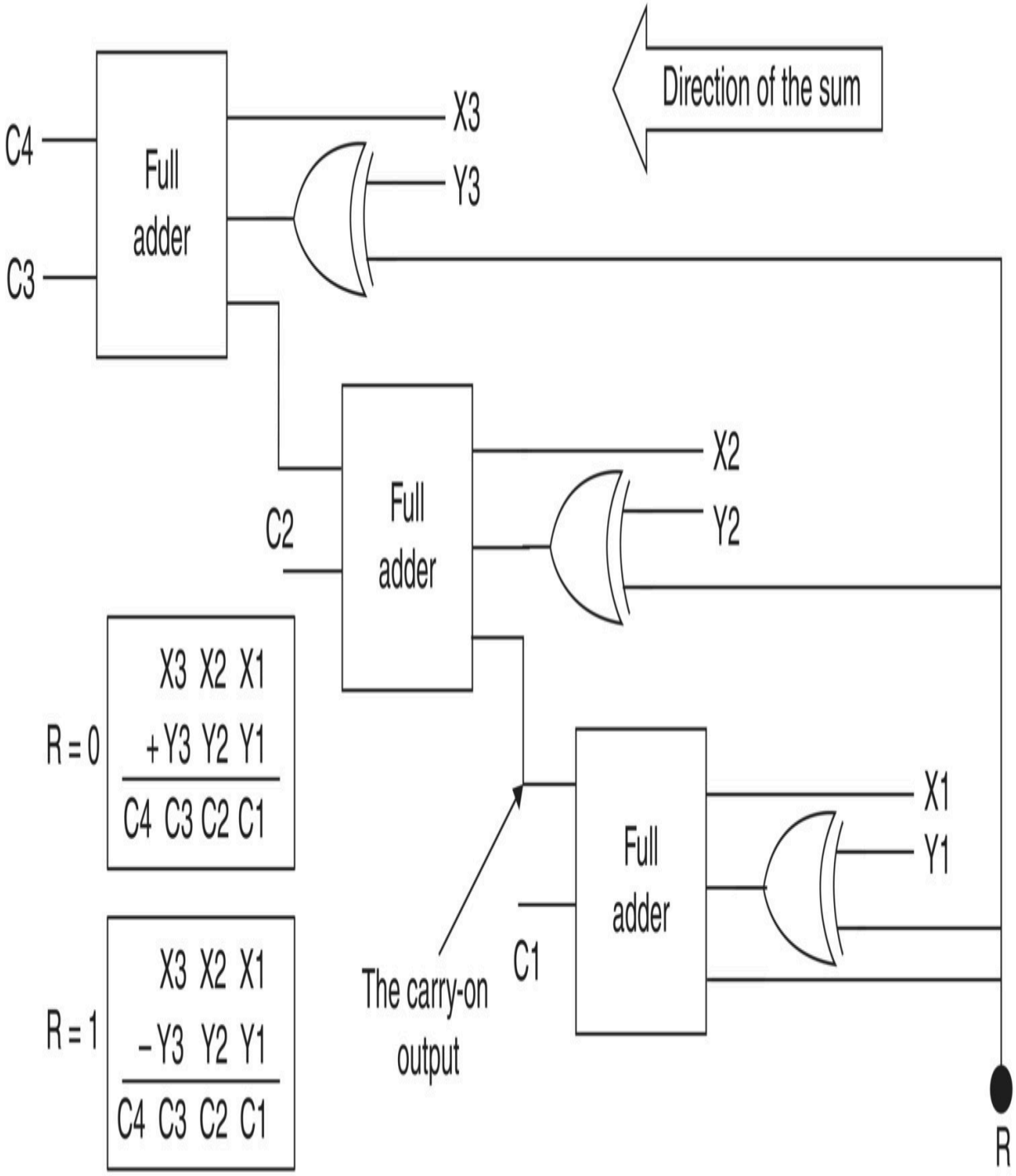


Figure 11.32 A full adder with the option to add or subtract the numbers depending if point R is 0 or 1, respectively.

Now that we know about subtracting by adding the complementary number let me show you another clever circuit.

Figure 11.32 is the same as the full adder in Figure 11.20 except that I have added three XOR modules between one number, the Ys, and the inputs of the full adder. Remember what the XOR module does (Figure 11.6). The output of the XOR is 0 if both inputs are the same, either both 0 or both 1. The output is 1 if the two inputs are different, one 0 and the other 1.

Suppose $R = 0$. If the Y numbers are 1, the output of the XOR is 1 and if the Y numbers are 0, the output is going to be 0, that is, if $R = 0$, the output of the XOR will be the same as the Y inputs. The XOR is transparent to the inputs and the full adders work exactly the same way as the adder in Figure 11.20.

Now suppose that $R = 1$. If any of the Y numbers is 1 the output of the XOR will be zero and if the Y numbers are 0, the output will be 1, that is, the XOR changes the value of the Y inputs, making them the complementary numbers. Furthermore, I connect point R to the carry-on input of the first adder. Remember from above that to get the complementary number I have to change the digits and add a 1. The circuit of Figure 11.32 does both addition if $R = 0$ and subtraction if $R = 1$.

Appendix 11.4 Dividing Digital Numbers

For completeness let me show you how we divide digital numbers. Suppose we want to calculate 53 divided by 4 (left of Figure 11.33). We look at the first digit of the dividend (5) and we look at how many times the divisor (4) goes into 5. Only once. So, we put 1 in the quotient and the remainder is 1. We bring down the 3 so now the 4 goes into 13 three times so the next number in the quotient is a 3 and $4 \times 3 = 12$, so we now subtract 12 from 13 getting a remainder of 1. The result of dividing 53 by 4 is 13 with a remainder of 1.

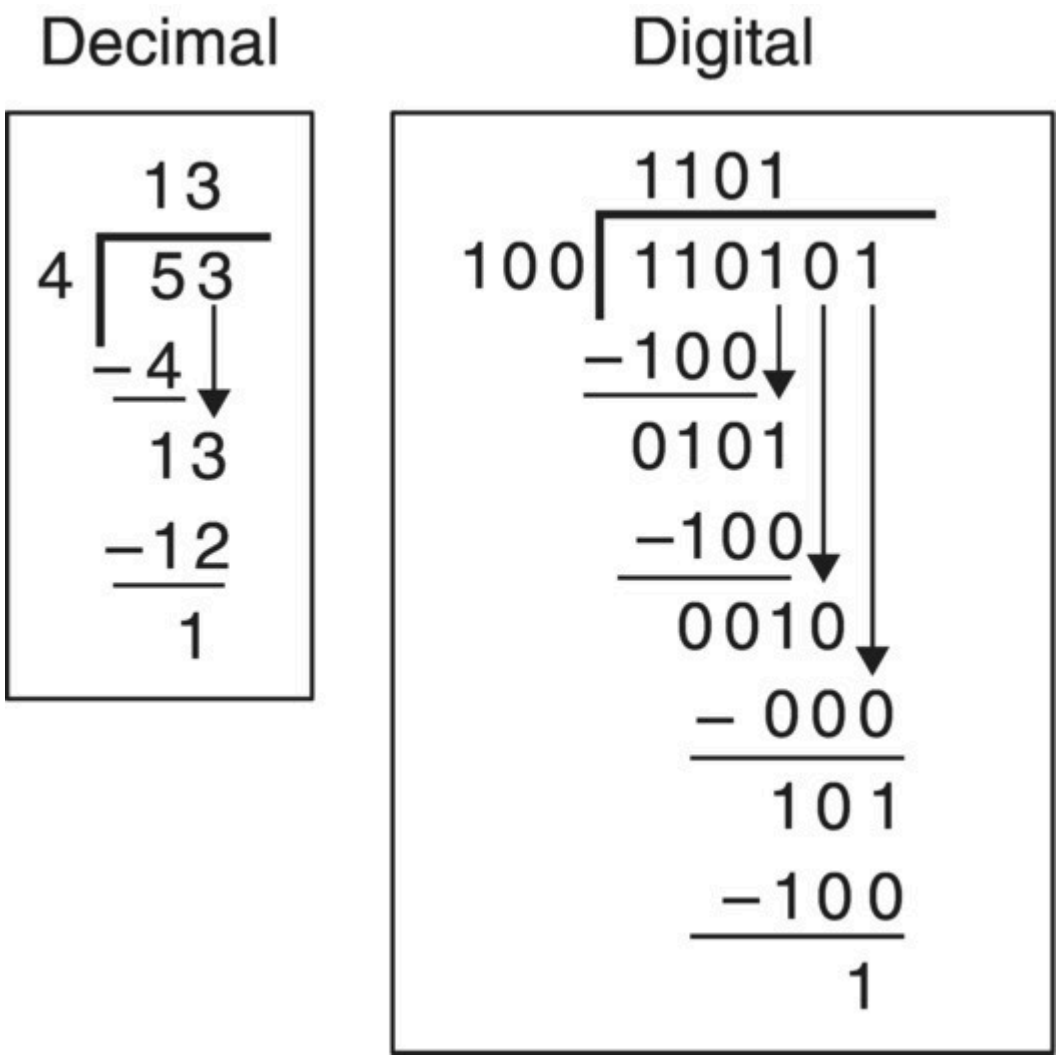


Figure 11.33 How we divide in the decimal (left) and the digital (right) systems.

We do the same thing with digital numbers. We look at how many times the divisor, 100 (4 in decimal) goes into the first three numbers of the dividend (110). It is only once, so we write a 1 in the quotient and subtract 100 from 110, giving a result of 010. We bring down the next number, a 1, and we have the same situation as before, so we subtract 100 from 101, getting a 1 in the quotient and a 1 in the remainder. Now we bring down the next digit, a 0, but now 100 is larger than 010, so we place a 0 in the quotient and subtract a 0 from the 010, getting, of course, the same number 10. When we bring down the last digit of the dividend, a 1, we get 101.

The quotient is 1, we subtract 100 from 101 and get 1. The result is therefore 1101 (13 in decimal) with a remainder of 1.

Appendix 11.5 The Author's Symbolic Logic Machine Using Relays

Just as an anecdote, Figure 11.34 shows me with a symbolic logic machine I built in 1962 using the relays and switches of an old pinball machine. This machine calculated the AND, OR, NOT, and IMPLICATION, the last very important in symbolic logic calculations, but one that is not used in mathematical calculations.

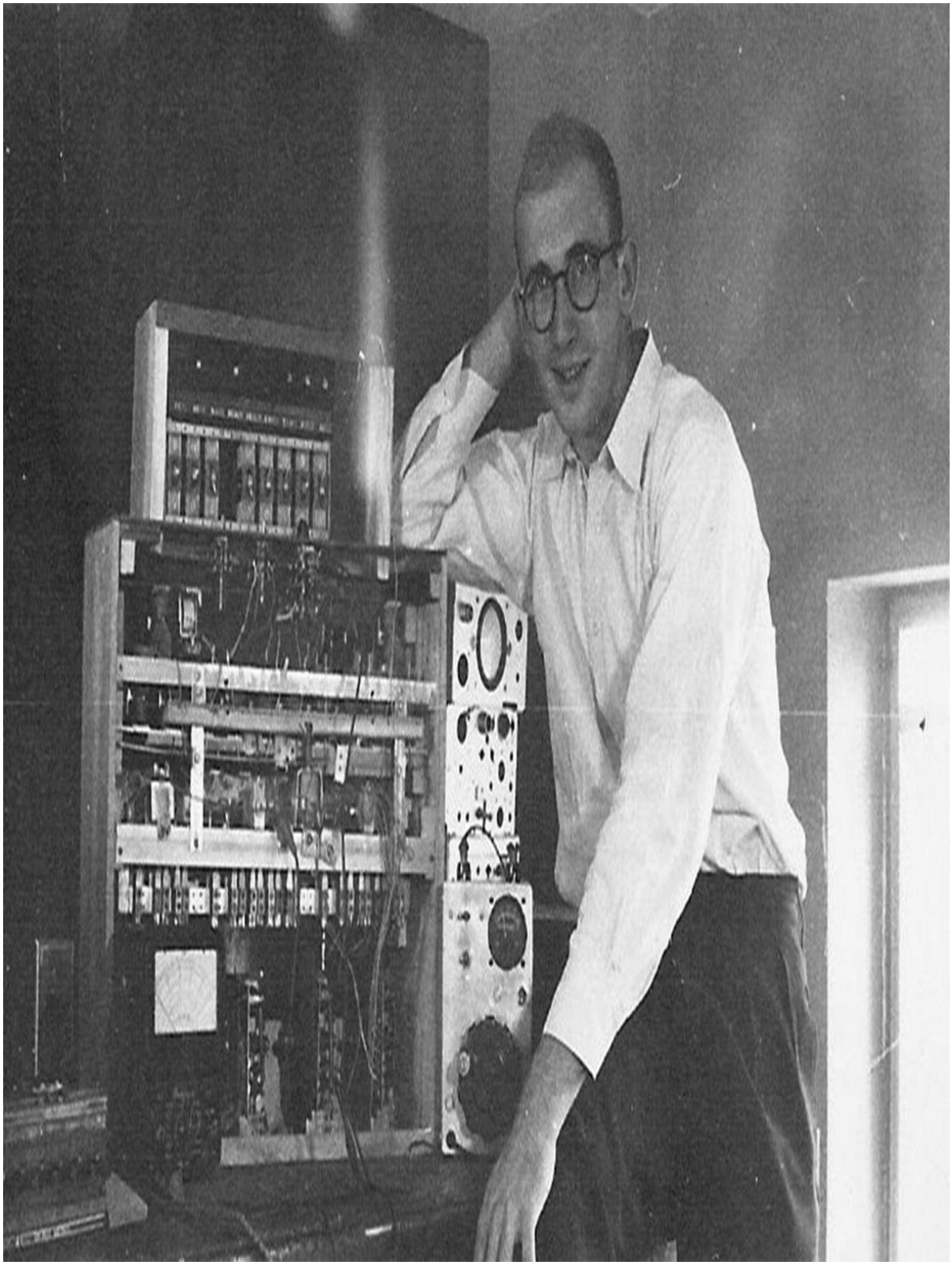


Figure 11.34 The author with a symbolic logic machine designed in 1962 using switches and relays from an old pinball machine.

12
VLSI Components

OBJECTIVES OF THIS CHAPTER

We are now ready to talk about more complex electronic system components that are an integral part of microprocessors, computers, cell phones, and many other devices that I discuss in the next couple of chapters. These include multiplexers that select signals from multiple inputs, demultiplexers that do the opposite, registers that store intermediate results, and all type of memories. We already have all the background we need to understand how these larger components work.

12.1 Multiplexers

A multiplexer is a component with many inputs and one output. We call it a MUX for short. It is an essential component in almost all large electronic systems.

The MUX is basically a selector switch, as I show in Figure 12.1.

The selector switch or the rotary switch on the left selects which of the four inputs I want to connect to the output, O. I show the schematic symbol of an electronic multiplexer on the right. In addition to the four inputs and the output, the symbol has two other inputs, the control inputs a and b, which determine the position of the switch arm and selects which of the inputs goes to the output.

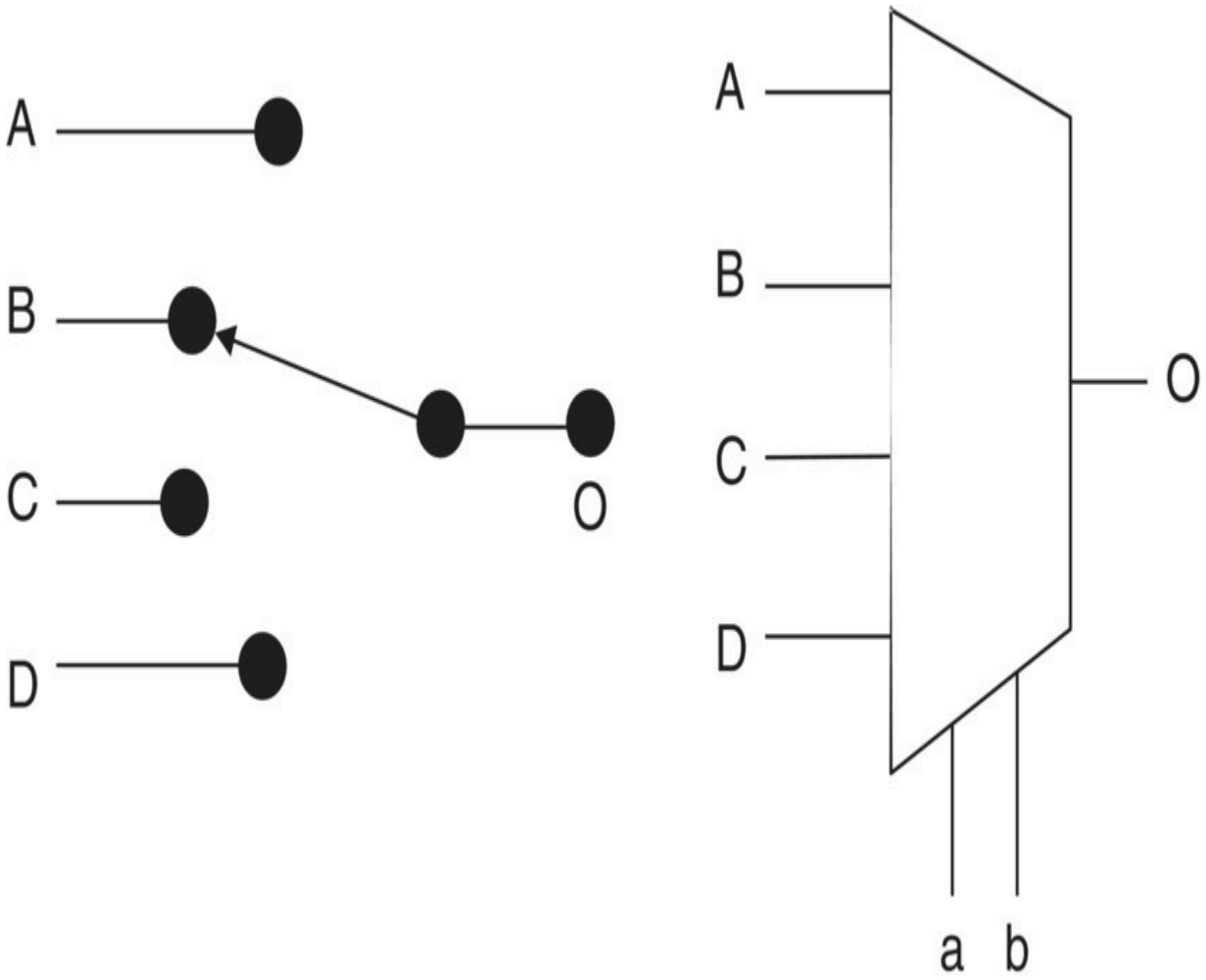
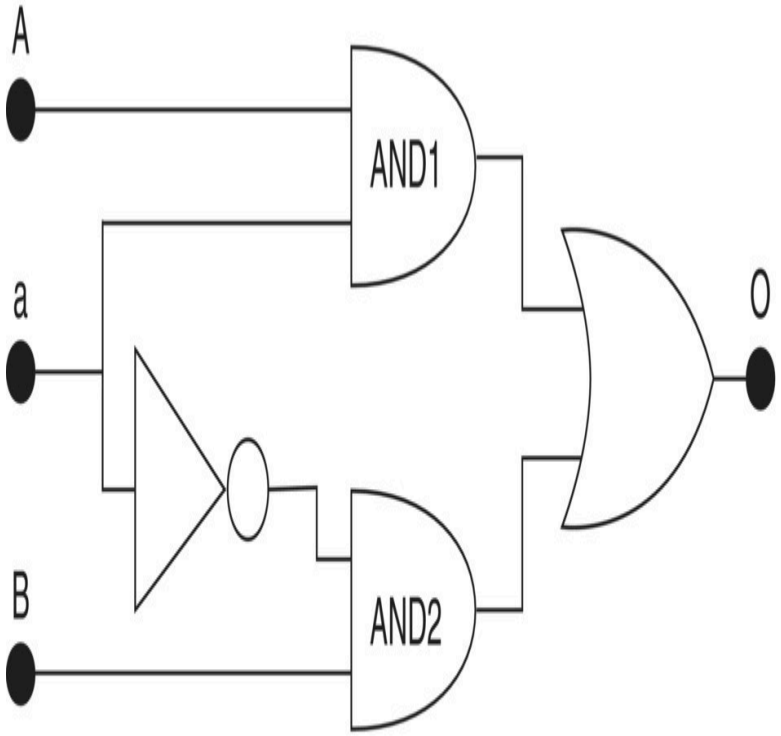


Figure 12.1 A MUX selects one of the many inputs, like a rotary switch. The symbol for a MUX is on the right.



a	A	B	O
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1

Figure 12.2 A 2 to 1 MUX implementation using two ANDs, one NOT, and one OR module with the truth table on the right.

The electronic circuit that performs the MUX function is not that complicated (Figure 12.2).

Recall that the output of an AND circuit is 1 only if the two inputs are 1, and 0 otherwise. Suppose now that I set the control line, a, to 1. Then one of the inputs of AND2 is *always* 0 and therefore the output of the AND2 is always zero no matter what the value of input B is. The output of AND1 is 1 only when input A is 1, and zero

otherwise. That is what the first four rows of the truth table tell us. The output, column O, is equal to the input A, no matter what the value of input B is. You can right away see that when the control line a is zero the opposite is true. Now the output column is equal to the B column. By changing the value of the control line, I can select which input I want to look at, A or B.

There are several other ways one can fabricate a MUX. A very common one is to use three NANDs and one NOT because this is easier to layout (see Appendix 12.1).

We can expand the MUX to four inputs (Figure 12.3). Notice now that we have four inputs, A, B, C, and D, and two control lines, a and b. Which input is connected to the output depends on the values of a and b, which can provide four combinations of 0s and 1s. If you look carefully you will see that each a–b combination turns ON just one of the AND modules. If both control lines a and b are 0, for example, only the uppermost AND module is ON. At least one of the inputs of the three ANDs below the first AND is 0, thus letting only the input A go through to the output.

We can fabricate an 8 to 1 MUX by adding four more AND modules and one extra control line. You can see the pattern. A 128 to 1 MUX requires seven control lines ($2^7 = 128$). For every control line I add, I double the number of inputs I can multiplex. The number of lines I can multiplex is 2^N where N is the number of control lines.

Often, though, we prefer to design larger MUXs using smaller modules. For example, an 8 to 1 MUX can be implemented with two smaller, 4 to 1, MUXs, as I show in Figure 12.4. We use a similar trick we have used before when we talked about arithmetic functions, such as the full adder (Section 11.9). We fabricated the full adder by combining two half adders and an OR cell. We use the same trick now.

If the control line c is 0, only the inputs from the upper 4×1 MUX are connected to the output, but if c is 1, only the values from the lower MUX go through to the output. The results of the truth table

for the 8 to 1 MUX are shown on the right of Figure 12.4. Notice though that we still need three control lines, a, b, and c.

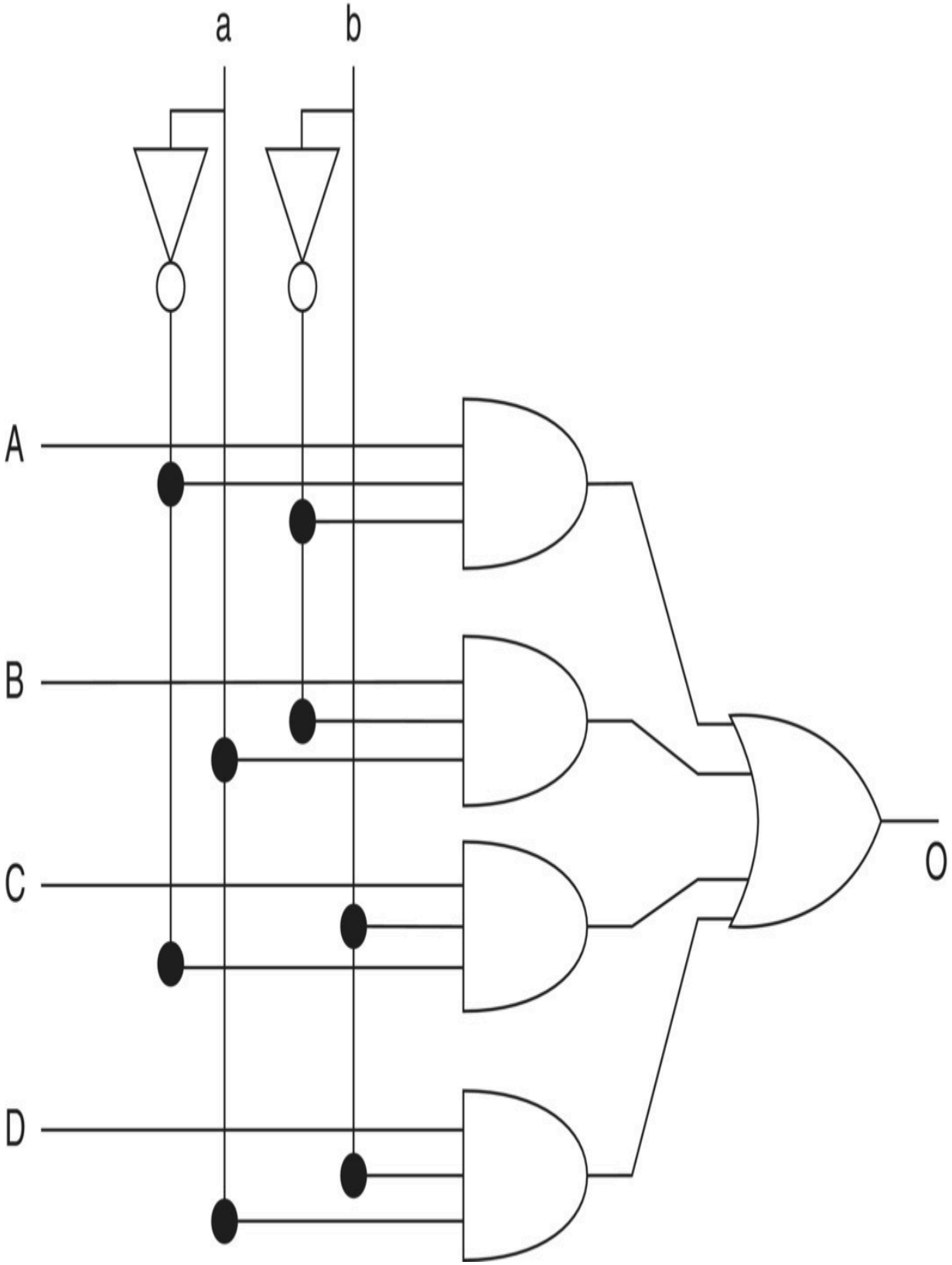
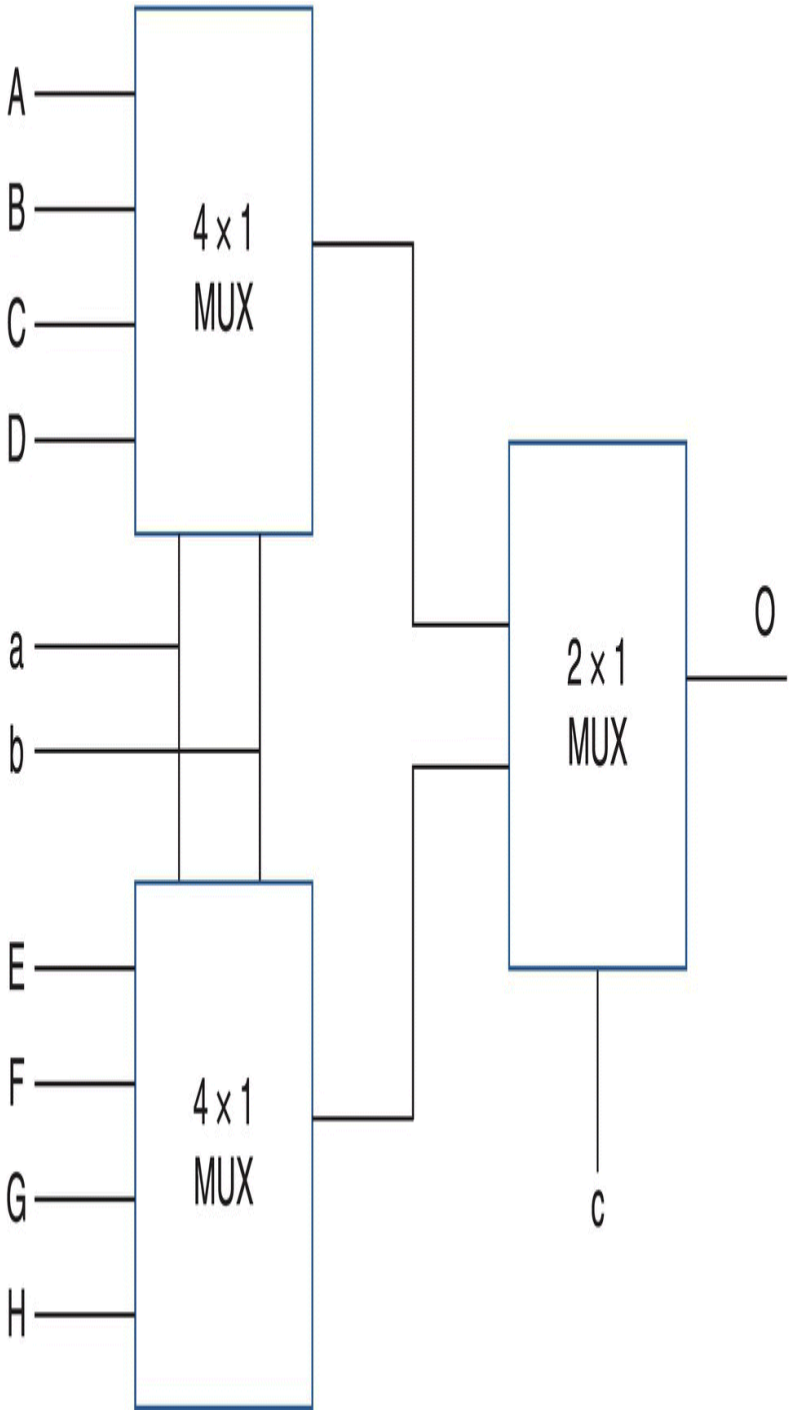

a
b
A
B
C
D
O

Figure 12.3 Implementation of a 4 to 1 MUX, using ANDs and NOTs. The two control lines result in four combinations and only one of the ANDs is ON at any one time.



c	b	a	o
0	0	0	A
0	0	1	B
0	1	0	C
0	1	1	D
1	0	0	E
1	0	1	F
1	1	0	G
1	1	1	H

Figure 12.4 An 8 to 1 MUX can be implemented by using smaller MUXs. Control line c determines which of the 4 to 1 MUXs is connected to the output. The truth table shows the results.

12.2 Demultiplexers

Demultiplexers (DEMUXs) do the opposite: they take one input and select which output we want to send the signal. The DEMUX are easier to construct than MUXs (Figure 12.5).

The AND modules are ON only when all the inputs are 1. If the input is zero, all the outputs will be zero no matter what the inputs are, but if the input is 1 only the AND module that has all 1s coming from the control lines will be ON. For example, suppose that a is 0 and b is 1. Only the third AND is ON and it passes the value of the input to the output C.

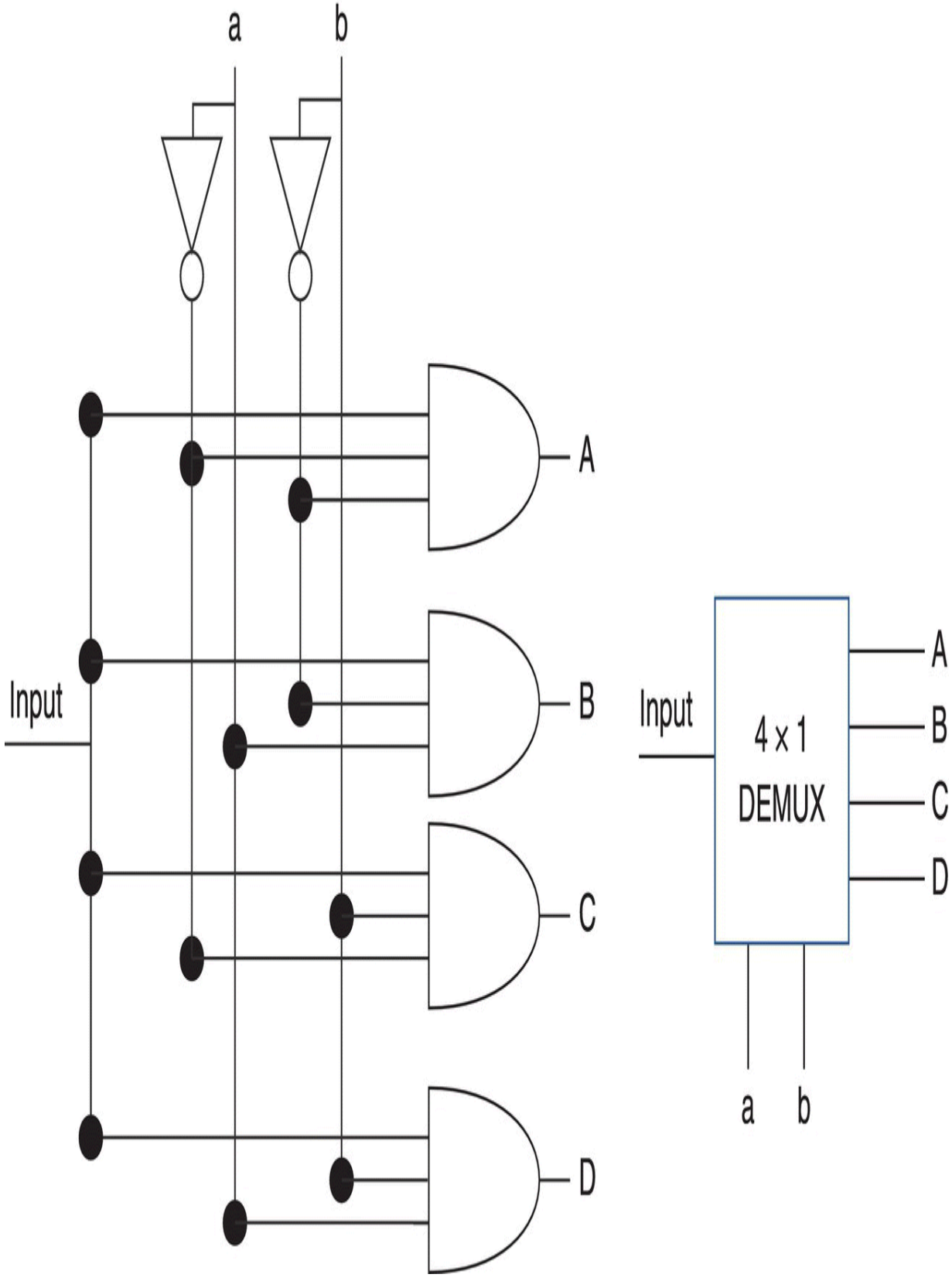

a
b
A
B
C
D
Input
Input
4 x 1
DEMUX
A
B
C
D
a
b

Figure 12.5 A 1 to 4 DEMUX using AND and NOT modules with the symbol on the right. The input is directed to just one of the four outputs depending on the status of the control lines a and b.

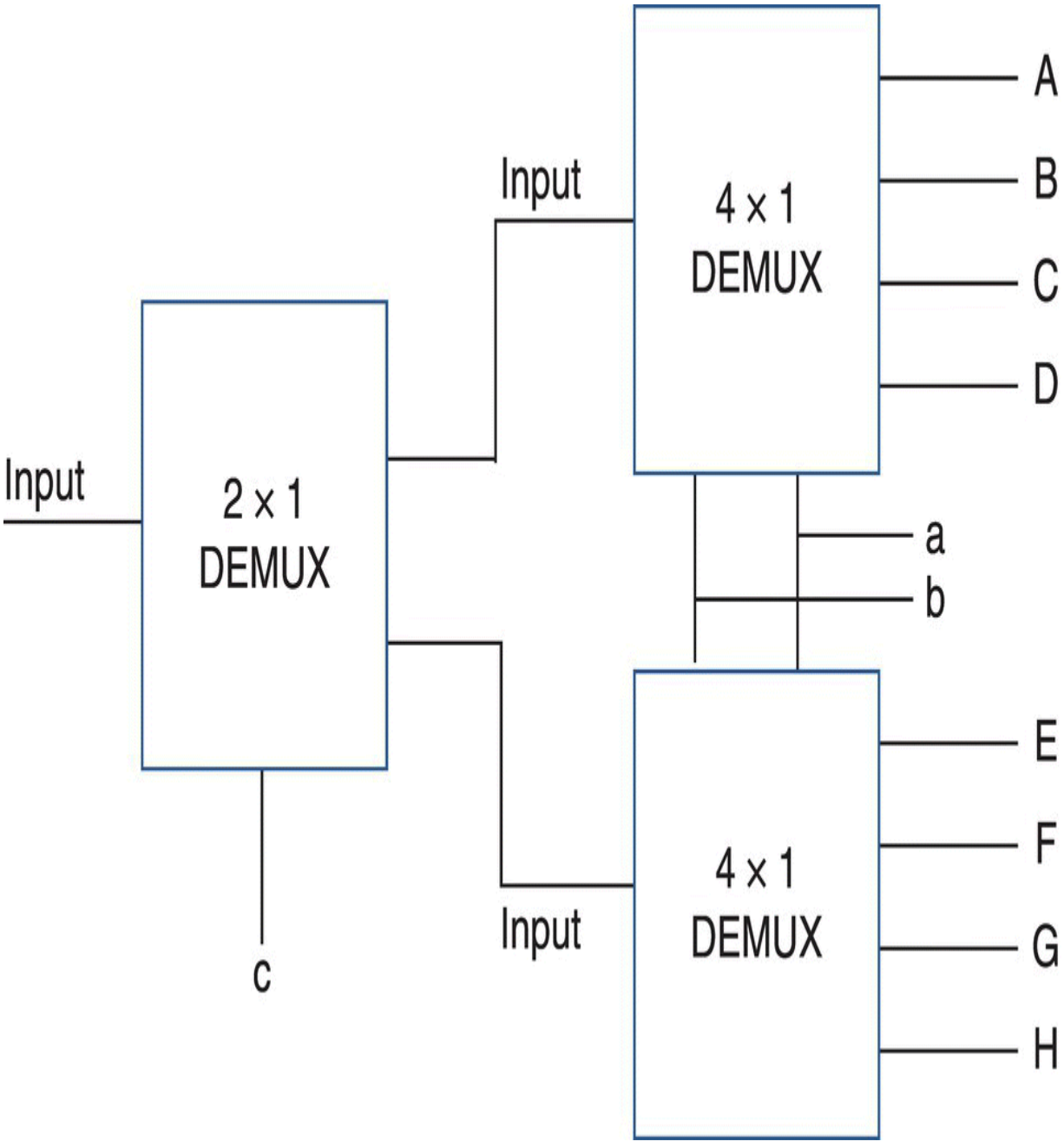


Figure 12.6 8 to 1 DEMUX constructed using smaller size DEMUXs.

As with the MUX we can expand this circuit either by adding more ANDs and control lines or, as we did with the MUX, use several of the smaller versions of the DEMUX, as I show in Figure 12.6.

12.3 Registers

A register is a circuit that stores digital words or numbers, actually a bunch of 1s and 0s. I covered flip-flops and latches in the previous chapter (Section 11.12). The flip-flop is basically a one-digit register. I set it up for 1 or 0 and it will stay this way until I decide to change it. A register, in its simplest form, consists of a bank of latches. Figure 12.7 shows this simple implementation of a register.

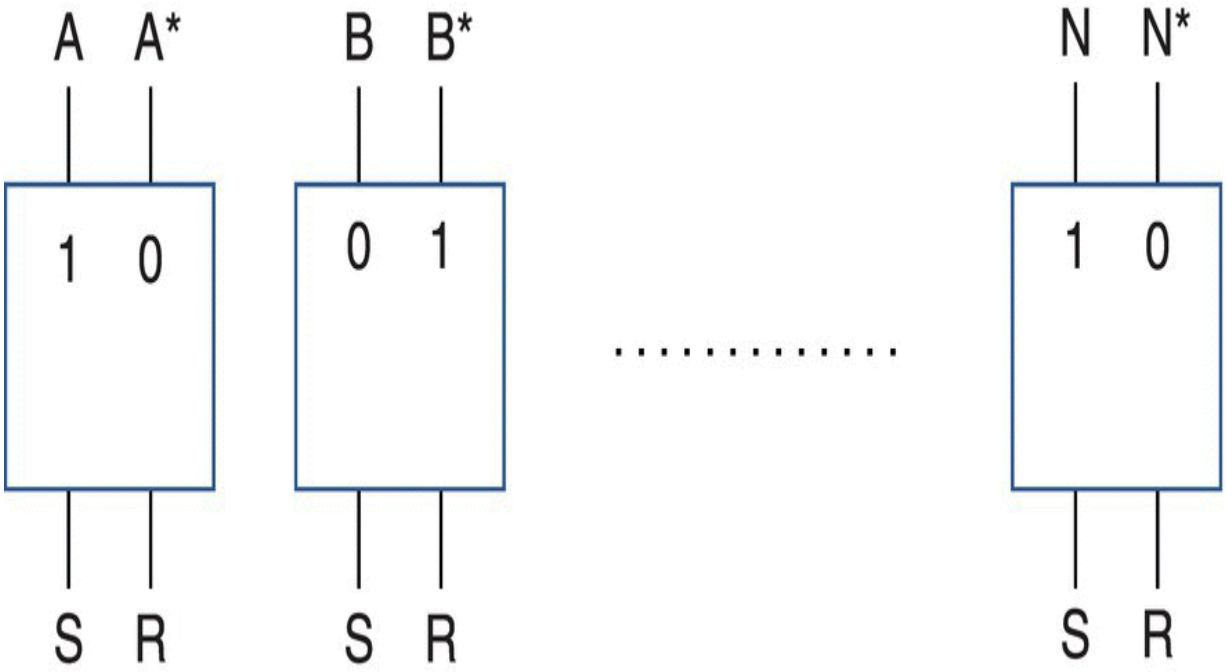


Figure 12.7 The register is composed of many latches with the non-asterisk outputs selected as the value of the latch.

If you recall (Figure 11.25), the latch has an input S, a reset R, and two outputs, one has a value of 0 and the other a value of 1. We can define the non-asterisked letter to be the master, so the value stored in the latches that define the status of the latch is the one without the asterisk. (This is arbitrary, we could use the other one, but we need to decide which is which.) The stored value in the first latch is 1, in the second latch is 0, and in the N-latch is 1. When we want to change its value we just send a signal to the set line.

What we want to do in most cases is transfer the value of one storage register to another, from the temporary store register to the

working register, for example. This has the advantage that we can take the working, second register, do whatever we need to do to the data, without destroying the original data, something like getting a duplicated photo, experiment with it until you like what you see, and never lose the original. I show this transfer in Figure 12.8.

I have connected the two registers with a bank of AND modules. The only time that the set values of register 2 are enabled and ready to read is when the common line "set" is 1. Then whatever is stored at the left of each latch of register 1 is duplicated at the set input of register 2. I show the simplest implementation. Notice the asterisked value of register 1 is not connected. In this transfer implementation, we first reset register 2 so that the master output is zero. Then we turn the set line to 1 and transfer the data from register 1. This requires two operations, first a reset followed by a set. There is another faster implementation of this data transfer that I show in Figure 12.9. I have added an extra AND module connecting the slave output (the one that has the asterisk) to the reset input. One input of all the ANDs is connected to the set line. When I turn ON the set line, both outputs of register 1 are transferred to register 2 and I do not need to reset register 2 before I transfer the data. Data transfer occurs constantly in microprocessor operations. This second implementation of the register transfer is twice as fast as the one with a single AND module per register unit. Yes, it is more complex to fabricate, it requires more real state, but it is twice as fast. Again, the designer has to choose: speed versus complexity and power dissipation.

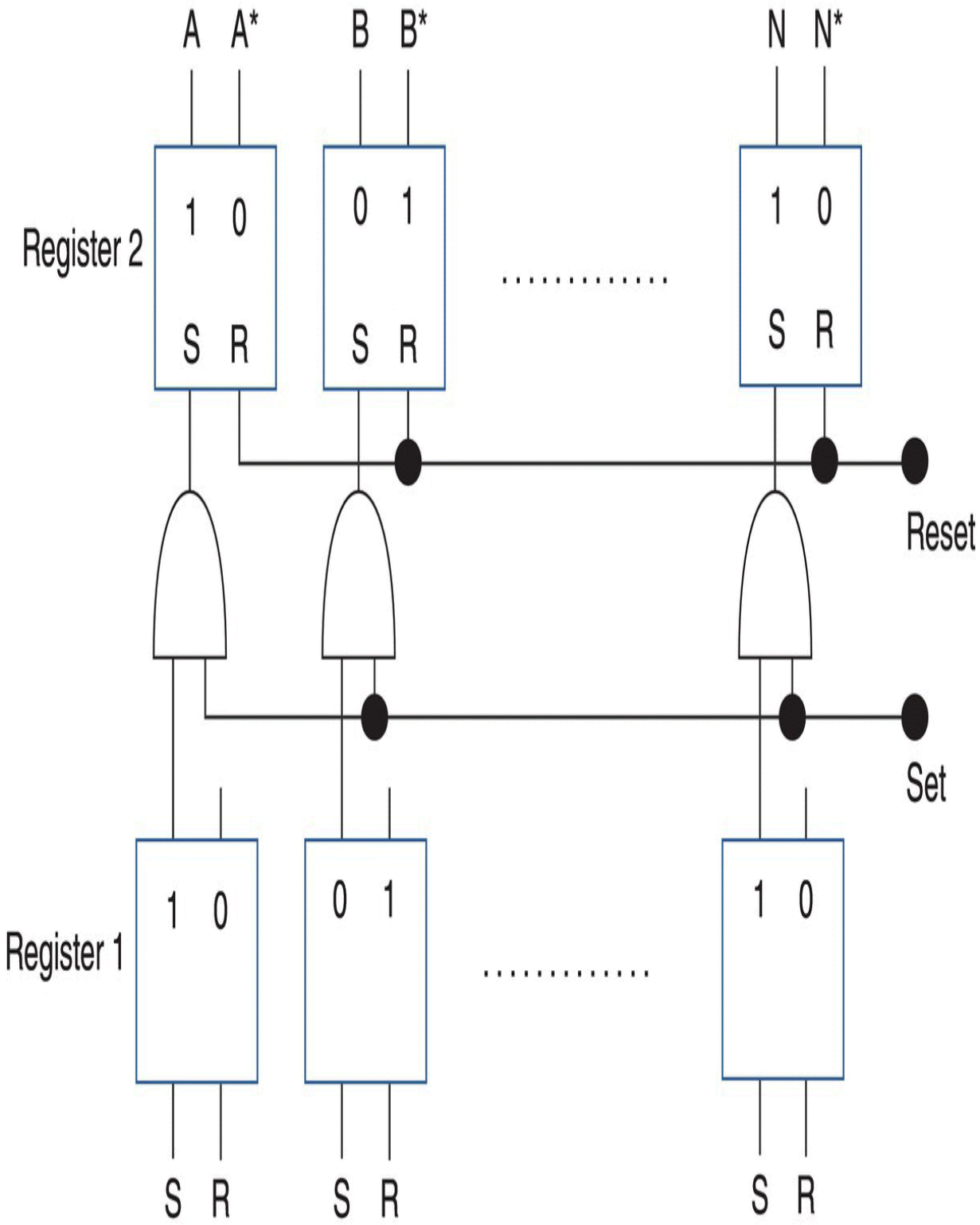

A A*
B B*
N N*
Register 2
1 0
S R
0 1
S R
1 0
S R
Reset
Set
Register 1
1 0
0 1
1 0
S R
S R
S R

Figure 12.8 To transfer data from register 1 to register 2, we turn ON the set control, S, turning the AND module ON and transferring the data. Note that we use only the A output of register 1. In some cases, we prefer to reset register 2 before we transfer the data.

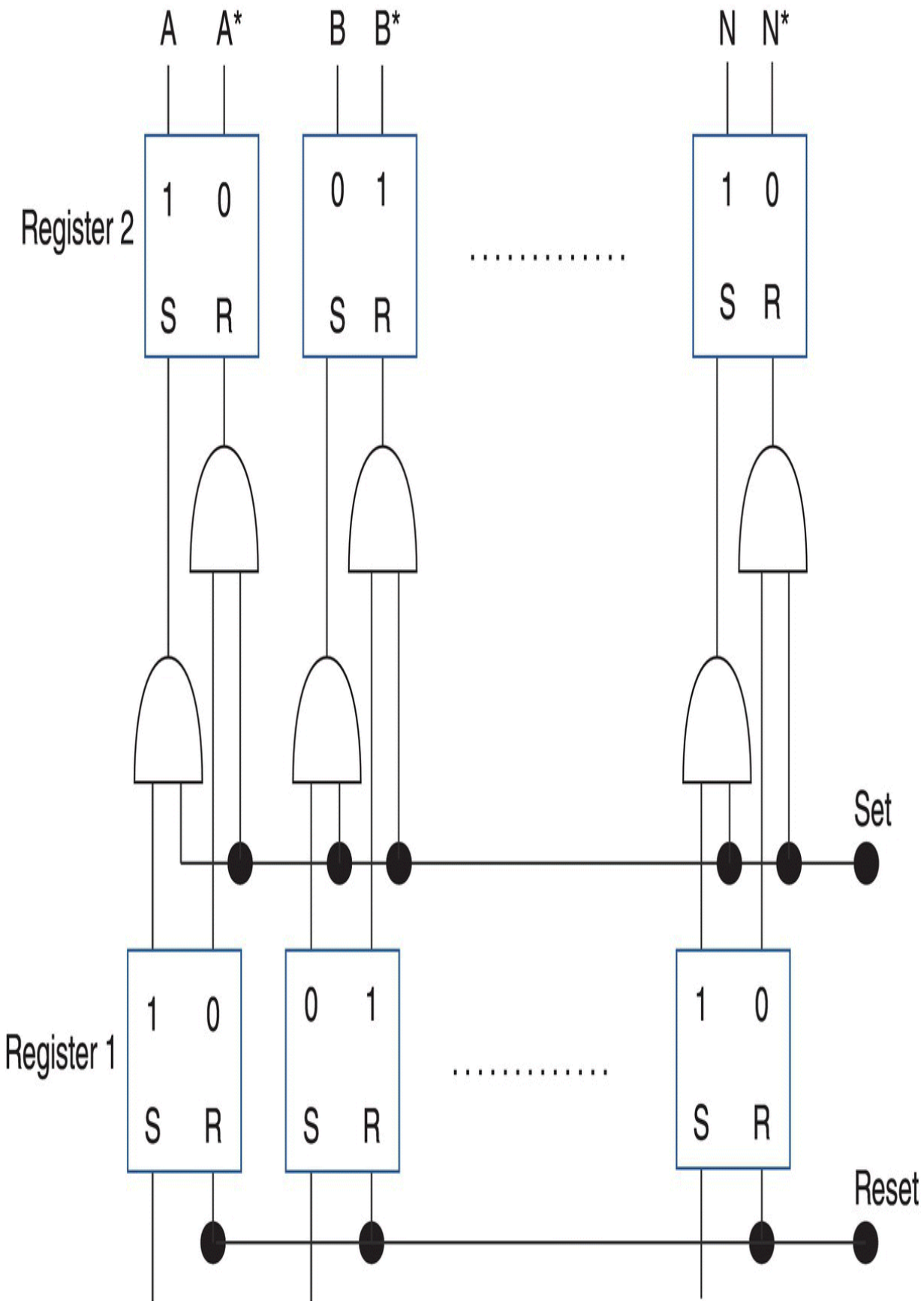

A
A*
B
B*
N
N*
Register 2
1
0
S
R
0
1
S
R
1
0
S
R
Set
Register 1
1
0
S
R
0
1
S
R
1
0
S
R
Reset

Figure 12.9 We can transfer the data faster from one register to another by adding another AND module.

12.4 Timing and Waveforms

I have not talked much about time or timing. In computing we have to do one operation after another after another. These sequential operations must be controlled and exactly timed. The timing system is like the conductor in an orchestra. Each microprocessor has a system clock. This clock is the metronome of the entire system. It just ticks at a certain frequency and generates a pulsed waveform. The conductor, following the metronome, decides when a particular orchestra section is supposed to come in, when they should stop, which ones play at the same time, which ones should be slowed down or speeded up. The timing system does exactly the same thing. Figure 12.10 shows the "partiture" that the electronic conductor, the programmer, uses to decide what to do next.

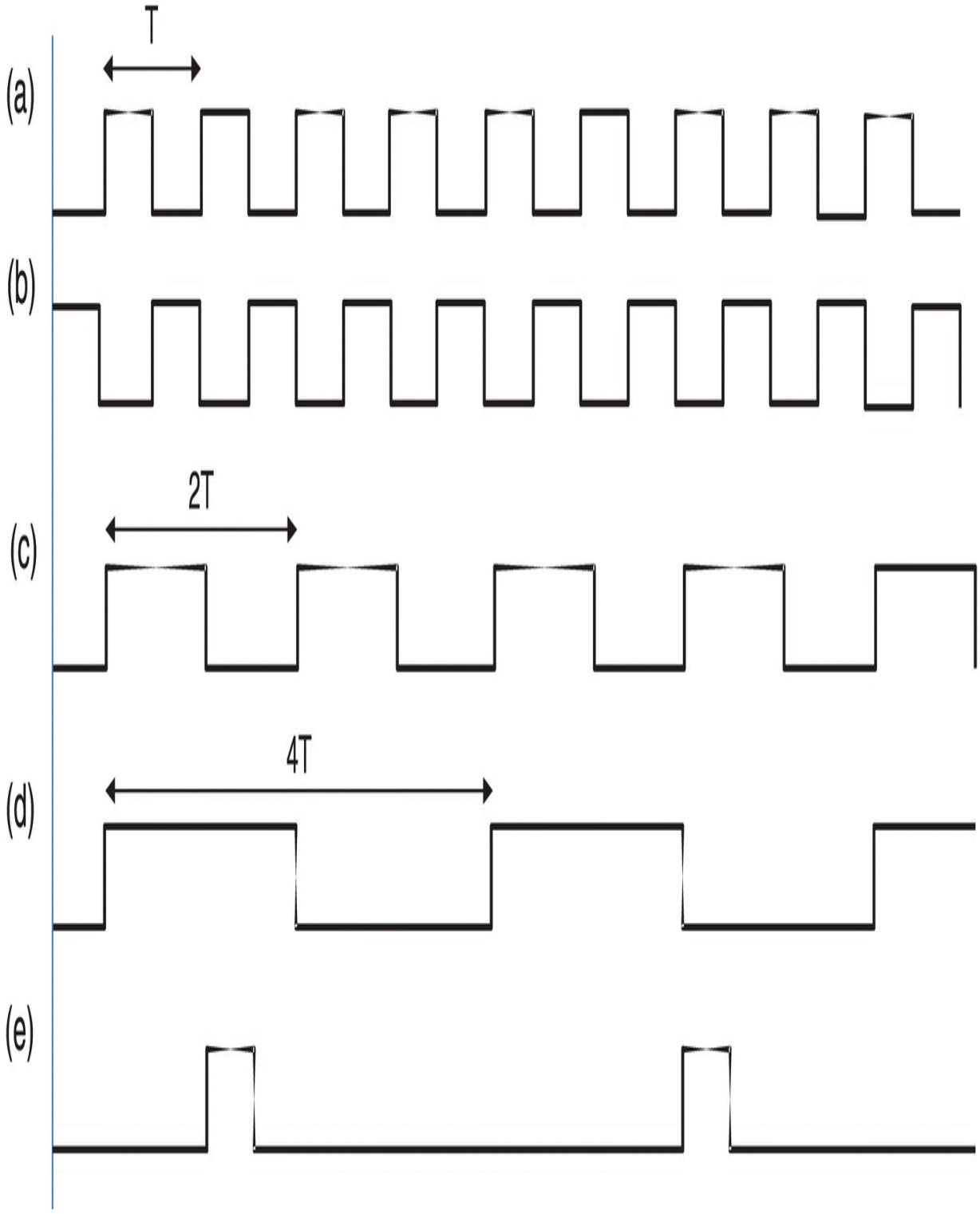


Figure 12.10 Many waveforms can be generated from the main system clock, the master clock (A). Different triggering schemes of latches allow us to change the timing of the waveforms.

All computers have an internal clock. It is a crystal oscillator which consists of a very thin quartz piece that precisely oscillates when we apply a voltage to it. The timing is all based on this internal clock that generates the set of pulses I show in Figure 12.10A. This is the metronome. Now, what we want to do is to create other timing schemes. The simplest one is when we want to change the polarity of the pulses, Figure 12.10B. This is very easy to accomplish; just add a NOT module and the 1s go to 0 and the 0s to 1. In other situations, we like to create pulses with longer periods, such as I show in Figure 12.10C and D. We accomplish this by using our ubiquitous flip-flops again. If we trigger the flip-flop status only when the master clock goes up, from 0 to 1, then the function generates a pulse twice as long as the original waveform, thus doubling its period. With the first rise of the master clock, the flip-flop changes from 0 to 1. Now the flip-flop has to wait a time T before the master clock goes up again from 0 to 1. Then it changes the value of the flip-flop from 1 back to 0. Using flip-flops on waveform C, we can create another waveform D, doubling again the period. If we want to increase the timing to $4T$, we just add another flip-flop stage.

Another possibility is that we want to generate one pulse but only after five master clock times, as I show in Figure 12.10E. Again, the flip-flops can help us. I have five flip-flops between the time I turn the pulse off to the time I turn it on. Then just one flip-flop to turn it back off. You get the idea. By cascading different combinations of the flip-flops, I can get any pulse train I need. And you thought that this circuit of two dogs biting each other's tails was just a curiosity!

There are different types of triggering. There is positive triggering when the clock is high or going high and negative triggering when the pulse is low or is going low. These circuits are also called "counters" because they are one way to count pulses.

One of the big problems with the counters is the delays between pulses. Suppose there is a small timing delay, t_D (Figure 12.11), between the master clock, top waveform, and the time that the

slave wave reacts and turns itself ON. This time can be very small, but as we generate different pulsed waveforms this time adds up until it is possible, as I show in the lowest waveform, that as the master clock goes up, the fourth slave waveform actually goes down, as you can see in Figure 12.11. This is not what we want and can cause a lot of asynchronous problems.

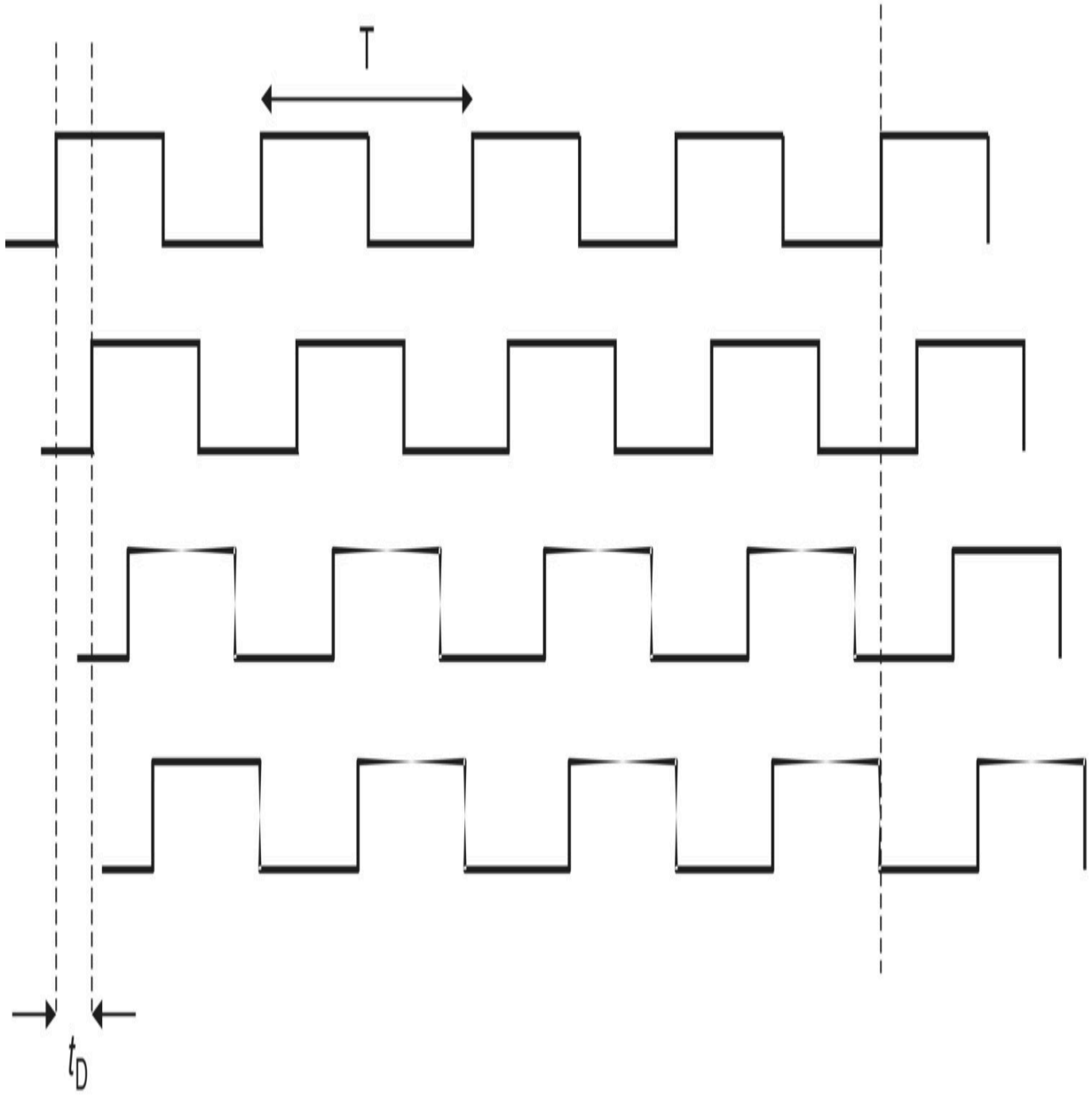


Figure 12.11 As waveforms move across the electronic system, there are timing delays that, after a few transfers, may result in a waveform doing the opposite of what it should.

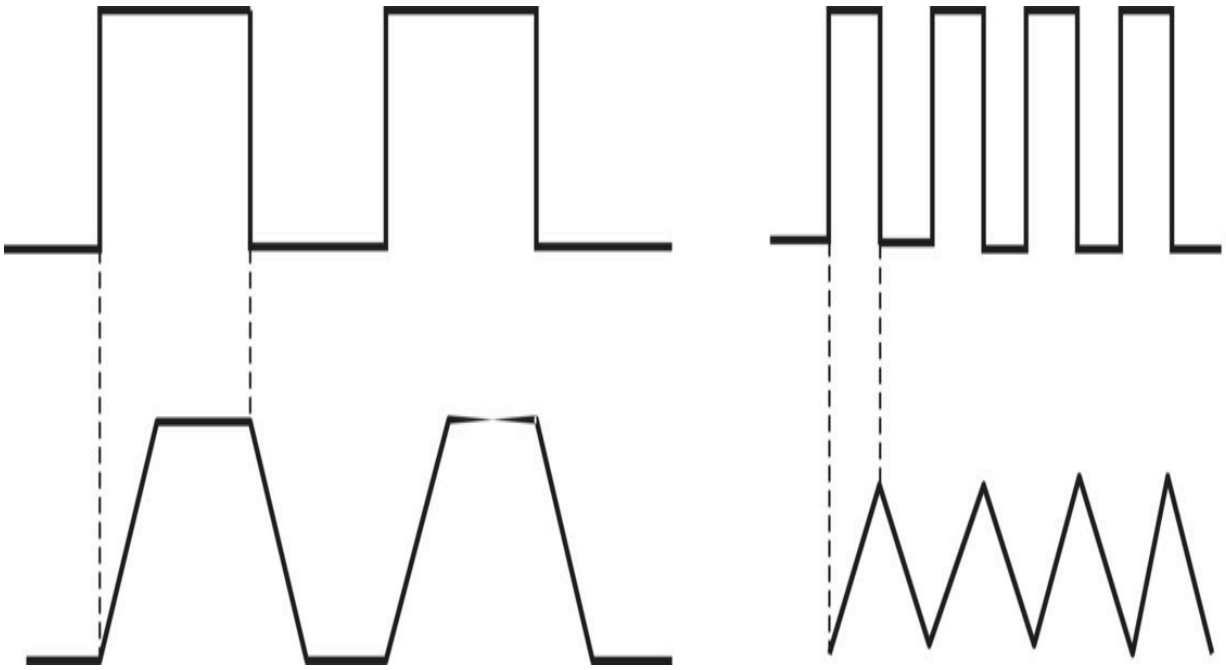


Figure 12.12 The rise and fall times of pulses limit the speed of the electronic devices.

Finally, another problem is that of distortion, which I addressed before. Figure 12.12 shows the problem. Pulses are not ideal, as I show at the bottom of Figure 12.12. It takes some time to reach the maximum value and time to go back to zero. These delays limit the speed of the system. The higher the speed, the faster the pulse has to go up and down. On the right I show what happens when the master wave increases its frequency. At some point the master wave will turn OFF before the slave has a chance to reach its top value. A lot of care has to be taken to optimize the operation of the system.

12.5 Memories

I have a photo stick with 64 Gb of memory. This photo stick, which is smaller and thinner than my finger, contains 64 000 000 000 cells. The chip inside is even smaller than the stick. This gives you an idea of the dimensions of each cell. If a photo has between 2 and 4 Mb, I can store up to 30 000 photos.

Memories are used in all electronic devices. They not only hold the data we store, but also the applications, the instructions, the location where we have stored something, even the procedures to transfer information from one place to the other. Therefore, the electronic memory chip, in addition to having cells that store the information, must have lines that power the cells and the electronics that allow us to select a cell and write and read the information we want or need. Not only that, but we should be able to read or write one cell at a time and do it quickly.

There are several types of memories. Some you can read or write in any order you want, these are called random access memories (RAMs). With other memories you have to scroll over a row or a column and read or write the information when the right cell becomes available. These are called sequential access memories (SAMs). Another set of memories is buffer memories. Memories are always slower than the central processing unit (CPU). Buffers are the memories we use to select the information we need to bring into the reach of the CPU. The cache, another type of memory, is similar to the buffer memory but is much faster so that the transition of the information to the CPU is as fast as the CPU needs it so no time is wasted.

The largest and slowest memory is like the city library and the buffer is like the shelf at home where I place the books I brought from the library. The cache is the night table or my desk where I keep a couple of books open at the relevant pages, so I can read or study them. It would very inconvenient if every time I wanted to consult some book I had to drive to the library, find the book, look at the information I need, and return the book to the library so that in the next hour, when I need another piece of information from the same book, I would again rush to the library, pick up the book, gather the information, return the book, and go back home. Even at home I have a place for the books I use, but when I am ready to do work, I want the book or books I am going to use now open on the desk ready for me to consult them and get the information I need at that moment. There are some other memories, sometimes called scratch-

pad memories. These are very close to the CPU and they are used for storing and retrieving intermediate CPU operations.

12.5.1 Static Random-access Memory

So, what is inside the static random-access memory (SRAM) unit cell? It consists of the module that we have already seen in Section 11.12, the flip-flop. Figure 12.13 shows a typical memory unit cell.

The two CMOS in the middle, M2 and M3, are the flip-flop, the two dogs chasing each other's tails. We already know that when A is 1, A* is 0, and vice versa, and it is stable, that is, it remains this way till we decide to change its value. The other two CMOS, M1 and M4, are just switches. There are five lines connected to this unit cell. These lines crisscross the entire memory array connecting all the cells in rows or columns that have the same function.

The three horizontal lines are the bias voltage, V_{CC}, the ground, and the word line. The first two horizontal lines bias the cell. The word line is connected to the gates of the CMOS switches, M1 and M4. When the word line is 0, both CMOS are OFF, and the flip-flops are isolated. If the word line is 1, then the two CMOS are ON and using either one of the two bit lines, b or b*, I can change the status of the flip-flops.

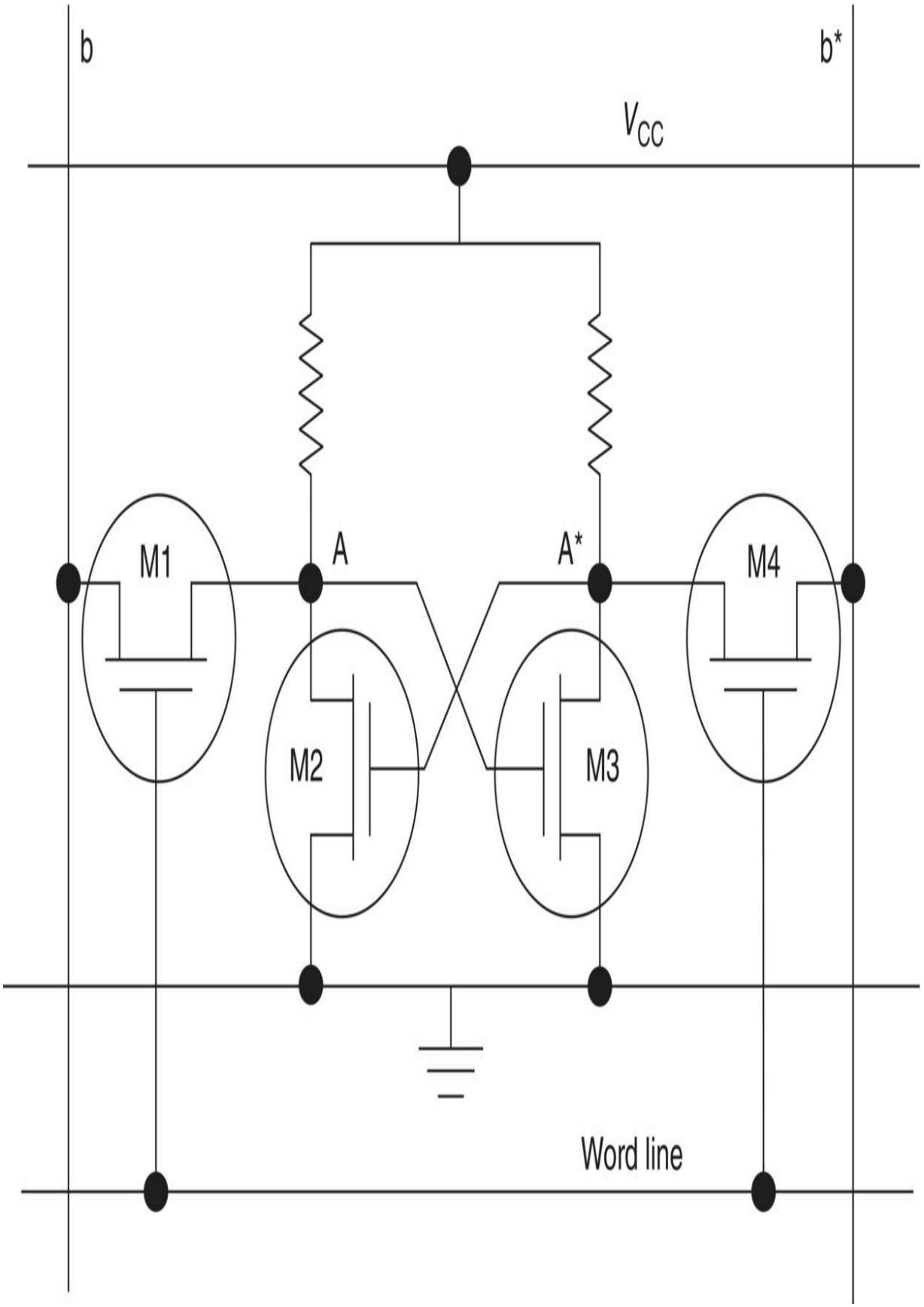

b
b*
V_{CC}
M1
A
A*
M4
M2
M3
Word line

Figure 12.13 A typical memory unit cell consists of a flip-flop in the center, the two CMOS, M2 and M3, and two CMOS switches on both sides, M1 and M4.

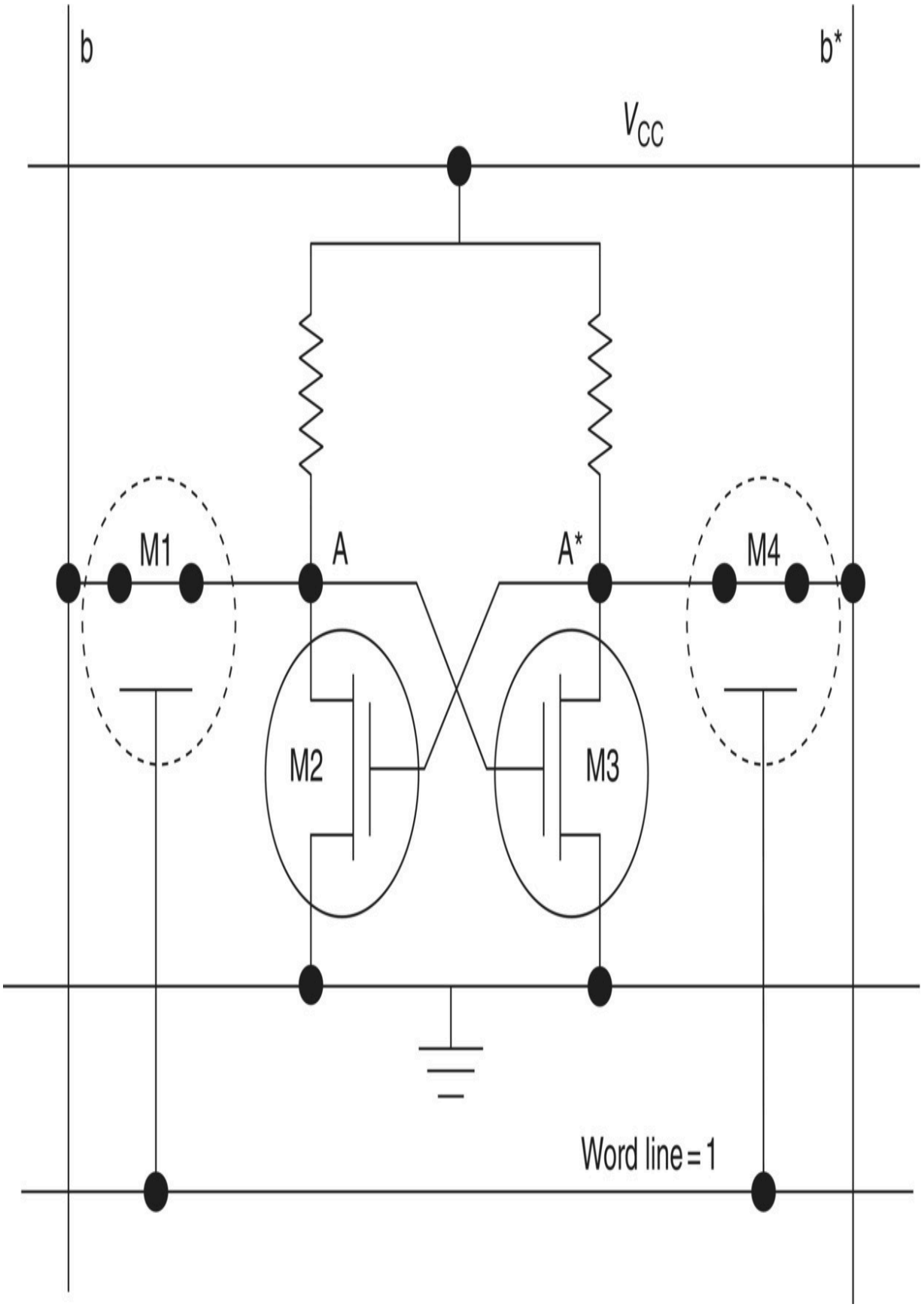

b
b*
V_{CC}
M1
A
A*
M4
M2
M3
Word line = 1

Figure 12.14 When the word line is 1, the CMOSs M1 and M4 are shorted, and the gates of M2 and M3 are connected to b and b*.

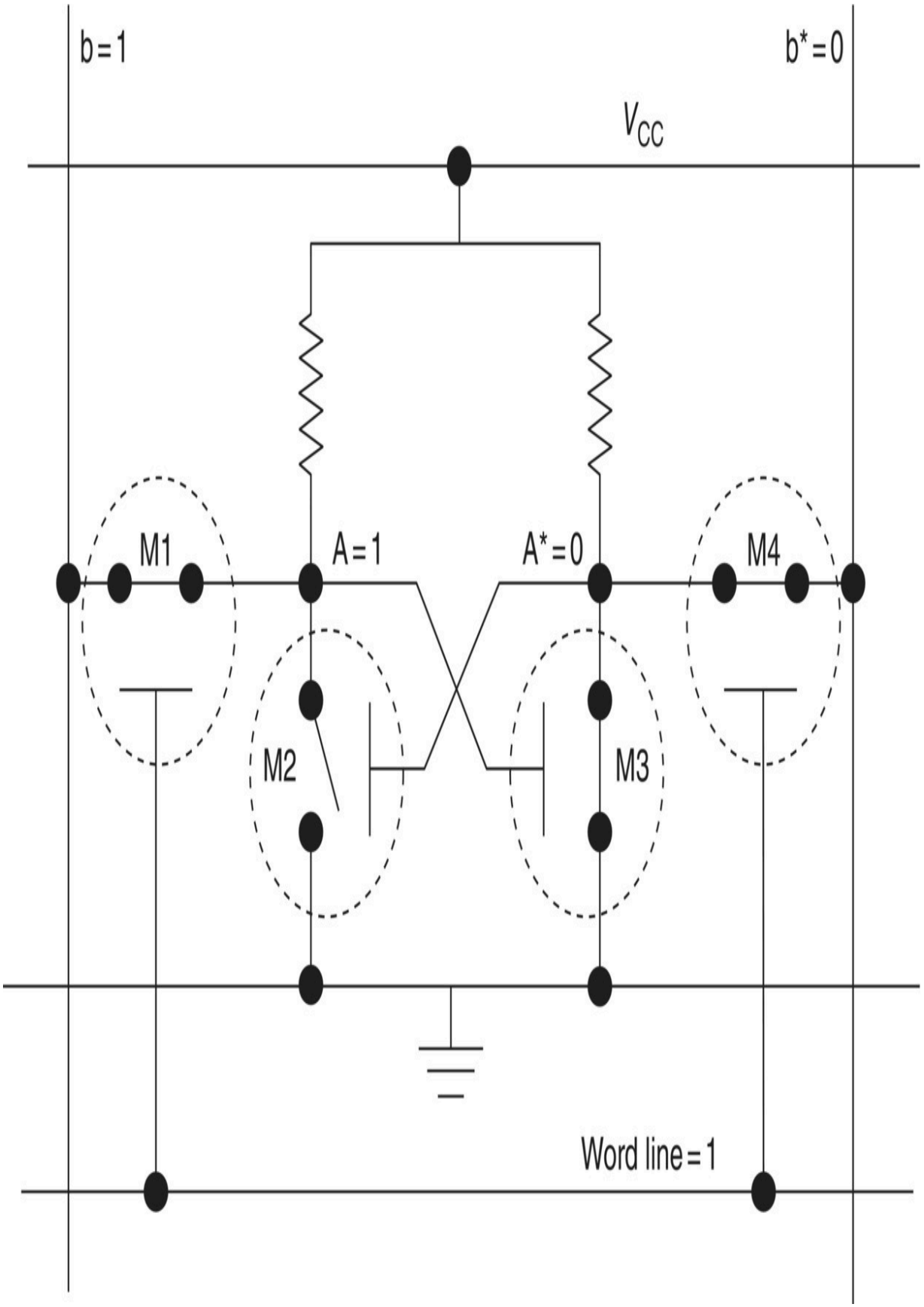

b = 1
b* = 0
V_{CC}
M1
A = 1
A* = 0
M4
M2
M3
Word line = 1

Figure 12.15 The CMOS in Figures 12.13 and 12.14 are replaced by switches. When word line is 1, A is connected to b, that is 1, and A* is shorted to ground and therefore is 0.

Let me explain the operation in a little more detail. Suppose that I turn the word line ON. Then M1 and M4 are shorted (Figure 12.14). I have replaced the CMOS M1 and M4 by shorts, so point A is connected to b and point A* is connected to b*. Now, let b = 1 and, therefore, b* = 0. The gate of CMOS M3 is 1 and therefore M3 is shorted and the gate of M2 is 0 and therefore open (Figure 12.15).

M3 is now shorted and M2 is open because A = 1 and A* = 0. You can see that point A is equal to b, which in turn is equal to 1. Since CMOS M2 is open, point A is not connected to ground and its voltage is equal to V_{CC}. On the right side the opposite occurs. Point A* is now shorted to b* and therefore it is 0. It is also 0 because it is shorted to the ground by M3. If I now change the word line from 1 to 0, M1 and M4 are open and A is going to be equal to 1 till I decide to change it again.

Now that we know how a unit cell works, we can construct an entire memory array (Figure 12.16). A memory consists of a matrix of unit cells, similar to the one we have discussed, which I show as blank square boxes. Each vertical array of unit cells has its own bit lines and each row has its own word line. Now, when I turn ON WL2, for example, the values of b, c, … x, y, and z are going to be written in the second row of memory cells. A vertical demultiplexer, which we have seen in Section 12.2, sequentially turns on each of the word lines. This is synchronized with the bit data so that the right data goes to the right cell. We use a MUX to read the cells in each row and we use a DEMUX to write on the cells.

The SRAM has fast access time and holds the information as long as it is connected the power supply.

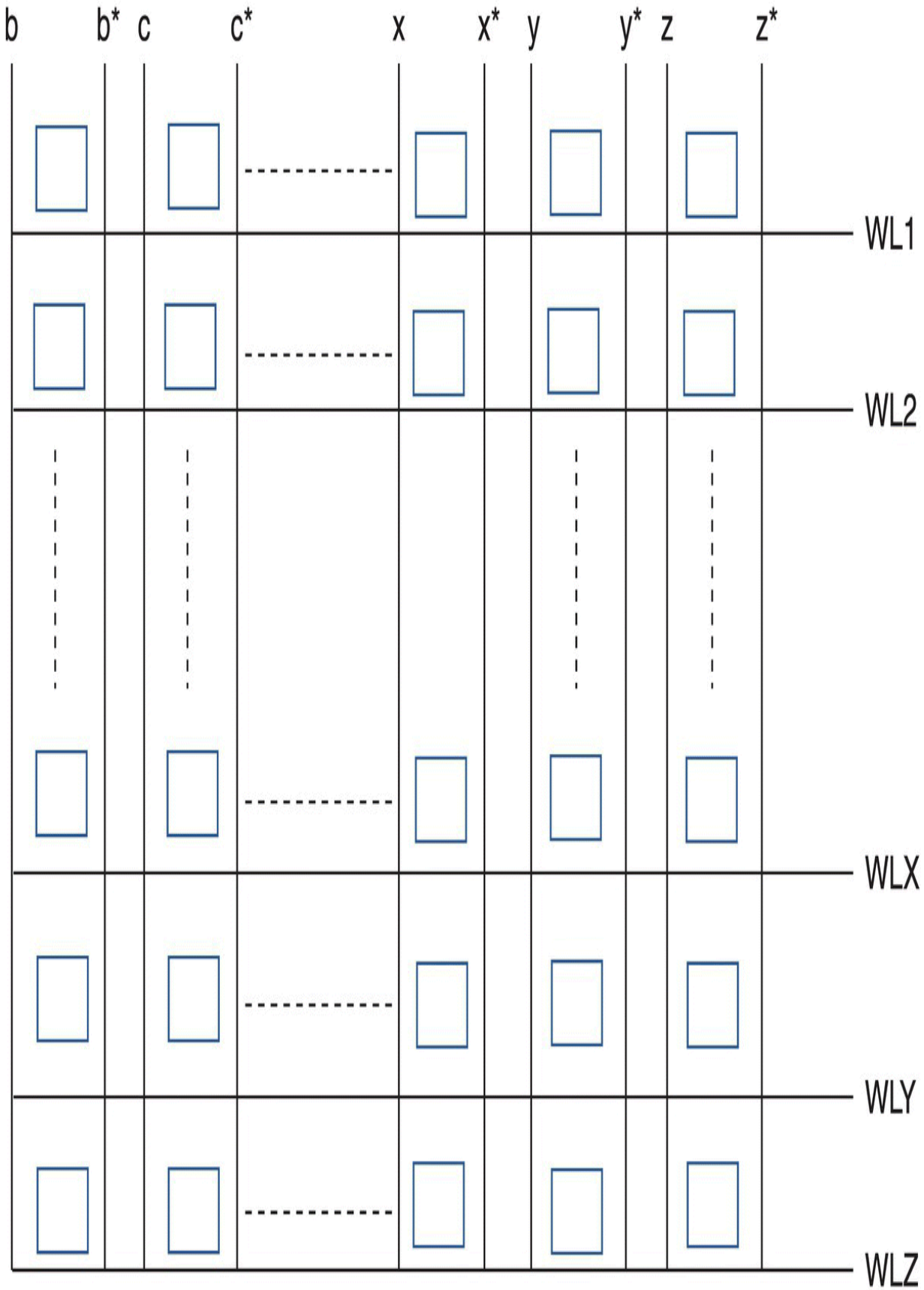

b
b*
c
c*
x
x*
y
y*
z
z*
WL1
WL2
WLX
WLY
WLZ

Figure 12.16 A memory chip architecture consists of a matrix of unit cells (the squares) addressed by bit lines on top (vertical lines) and word control lines on the sides (horizontal lines). I do not show the voltage supply or the ground lines.

12.5.2 Dynamic Random-access Memory

Dynamic random-access memory (DRAM) has a very simple unit cell, very compact so you may have many more cells per unit area. It has two problems: first it is slower than SRAM and second it needs to be refreshed constantly. Additionally, the refresh function always has preference over any other operation the computer may need. We do not want to lose the information. Typically, we need to refresh the memory every few milliseconds. Figure 12.17 shows a small array with the unit cells and the operating lines. First let's talk about the unit cell. It consists of just one capacitor and one MOSFET that acts like a switch. Notice also that each row of cells has just one input line, a, b, c . . . and 1 to Z word lines. So, in addition to having a much smaller unit cell than the SRAM, it also has two fewer lines crisscrossing the mosaic, no power supply, and no inverse input lines. How it works is easy. We turn ON, one word line at a time, all the FETs on that specific vertical line, that is shorted, and the capacitors in that vertical line charge to whatever the value of the input lines is. For example, if I turn ON WL3, that is, I make WL3 = 1 and the cell inputs a and c are 1 and cell b is 0, the three capacitors on the third line, from the top down will have values of 1, 0, and 1.

If unit cells are so simple, why don't we use them in every memory chip? The problem is that the MOSFETs are not perfect switches, that is, the resistance does not go from zero to infinity. We talked before about leakage currents. When I charge a capacitor, the voltage of the capacitor is equal to the voltage of the input gate (Figure 12.18), but as time goes on the capacitor discharges through the not-ideal switch.

Instead of just charging the capacitor to a voltage equivalent to what we consider a bit = 1, which I show in Figure 12.18 as V_1, we charge the capacitor to a higher voltage, V_{CC}. When we turn the FET OFF, the charge on the capacitor starts leaking out, thus decreasing and therefore the voltage across the capacitor decreases (remember in a capacitor $V = Q/C$, Eq. (6.14)). At some time later, t_1, the voltage across the capacitor will be lower than the voltage I consider to be 1. It is not zero, but it is somewhere in between, and it will confuse the system (is it 1 or is it 0?). In digital logic we have either a one or a zero, a three-quarters value is meaningless. So well before t_1 we need to refresh the value and set the specific capacitor back to V_{CC}, so, as in all engineering designs, using large capacitors increases the holding time while smaller capacitors allow more cells to be fitted into the chip area.

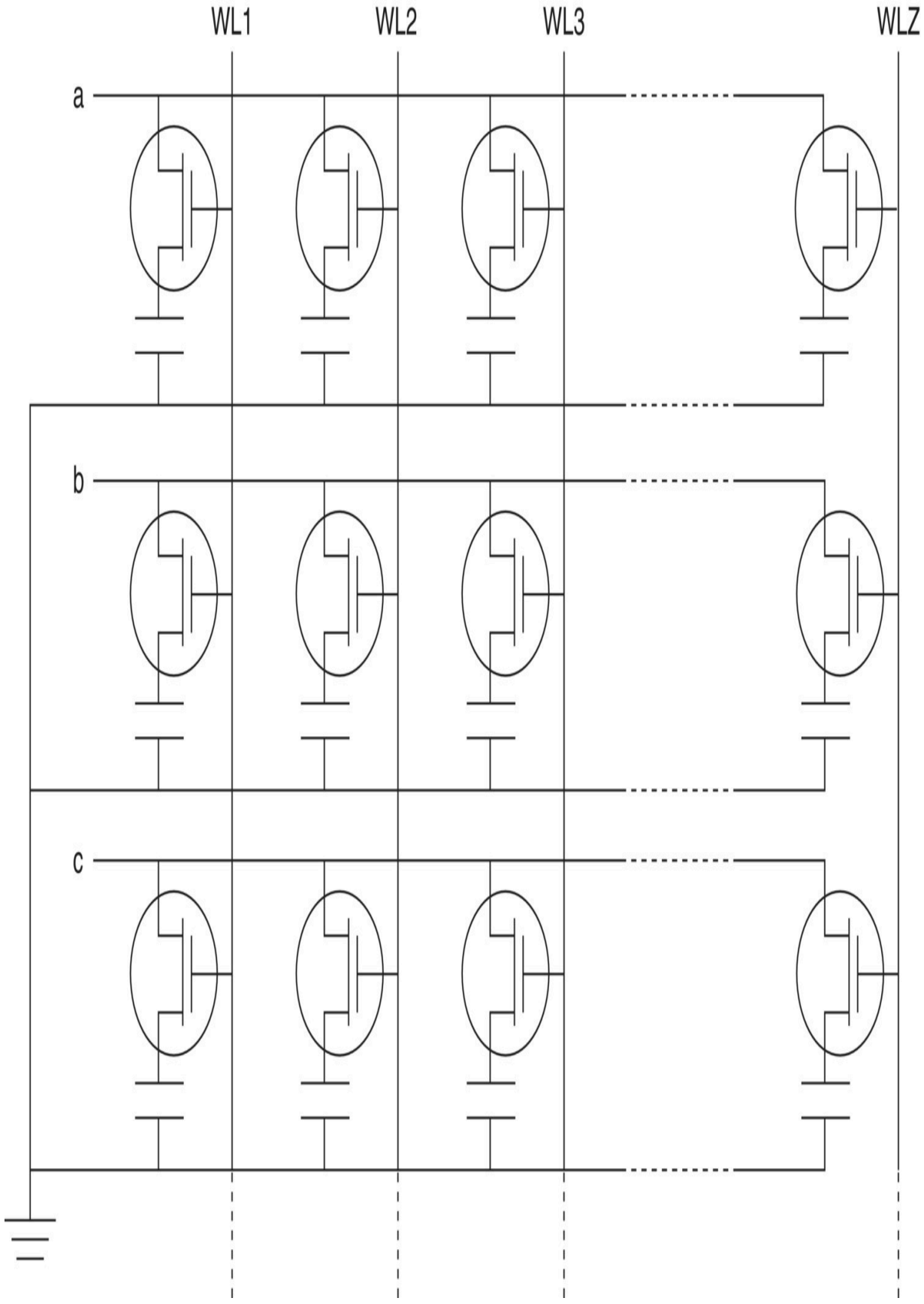

WL1
WL2
WL3
WLZ
a
b
c

Figure 12.17 The array of DRAM cells is addressed by a single input line (horizontal) and a word line (vertical) per column and ground.

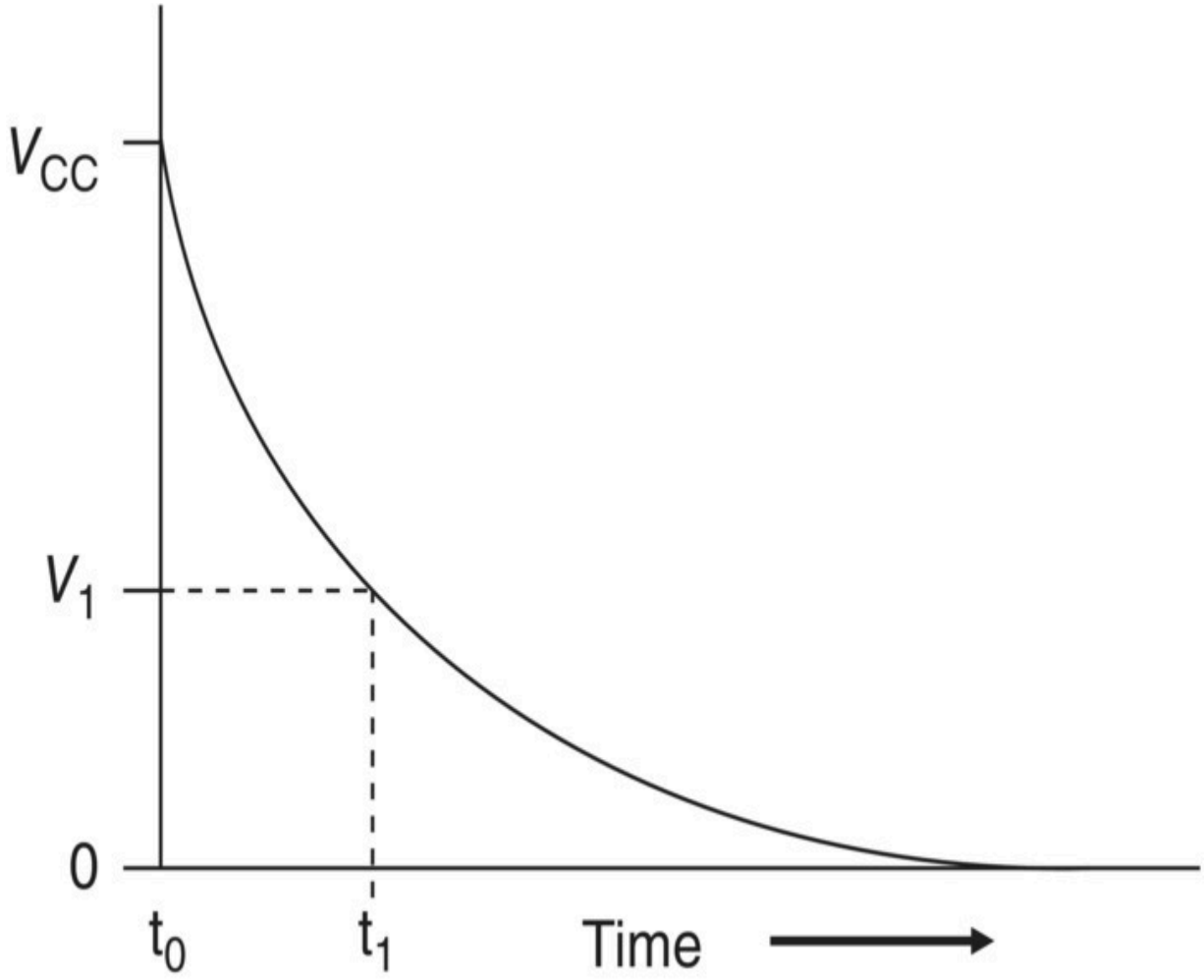


Figure 12.18 The capacitor charges initially to the full voltage, V_{CC}, but it discharges slowly as a function of time. After a time t_1, the capacitor goes below the voltage required to call it 1.

To refresh the value any time before t_1 we read the value, transfer this value to another capacitor, amplify its voltage, and write the information back again into the capacitor. You can see that even though they are very small, and the number of unit cells can be very large, the peripheral electronics needed to refresh their value is more complex, requiring specialized timing schedules and extra power supplies. On the positive side, reading and writing are faster

and use less power. In any electronic design there is never a free lunch.

12.5.3 Read-only Memory

As the name very clearly implies, read-only memories (ROMs) are fabricated so that the information is permanently stored in the cells and cannot be changed. There are many applications where there is no need to change the content of the memory, for example tables and apps. I do not want a glitch or any of my typing errors changing their values. I show in Figure 12.19 a way to do this.

First let's consider the unit cell. It consists of only one CMOS FET. All the sources are connected to ground and all the drains are connected to one of the vertical bit lines, b, c, or d. The gates are connected to one of the horizontal word lines. When one of the word lines is ON, all the CMOS on that line are ON and thus shorted. That means that the collectors of the CMOS of that line are connected to ground. The CMOS on the other word lines are OFF and thus open, and their collectors are connected to the bias V_{CC}. Let me give you an example (Figure 12.20). Suppose that the word lines WL1 and WL3 are OFF and WL2 is ON, as shown in Figure 12.20. Notice that the only bit line that is not shorted to ground is line b, all the others, shown in bold lines, are shorted to ground. Since line b is not shorted, there is no current through the b line and therefore no voltage drop on the resistor and the whole b line is equal to V_{CC}. The other lines, c and d, are shorted to ground and the voltage V_{CC} drops across the resistors. Therefore, the status of the lines b, c, and d are 1, 0, 0, respectively (or number 8 in the decimal system). If instead of turning WL2 OFF I were to turn OFF WL1 or WL3, I would get for WL1 ON 0,1,0 (number 2) and 1,1,0 (number 10) for WL3 ON.

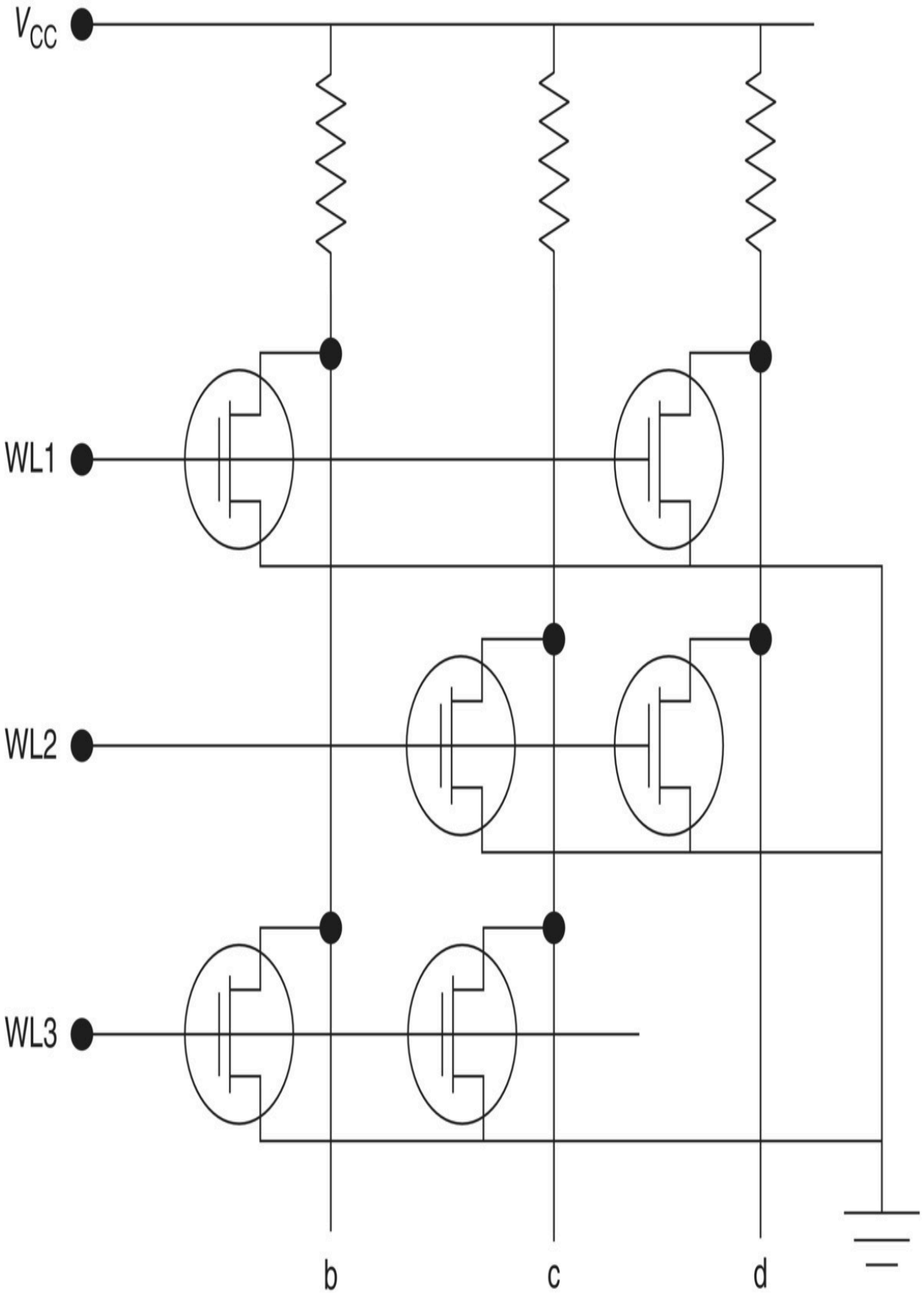

V_{CC}
WL1
WL2
WL3
b
c
d

Figure 12.19 A ROM consists of CMOS arranged in such a way as to ensure that only one bit line has all the CMOS OFF.

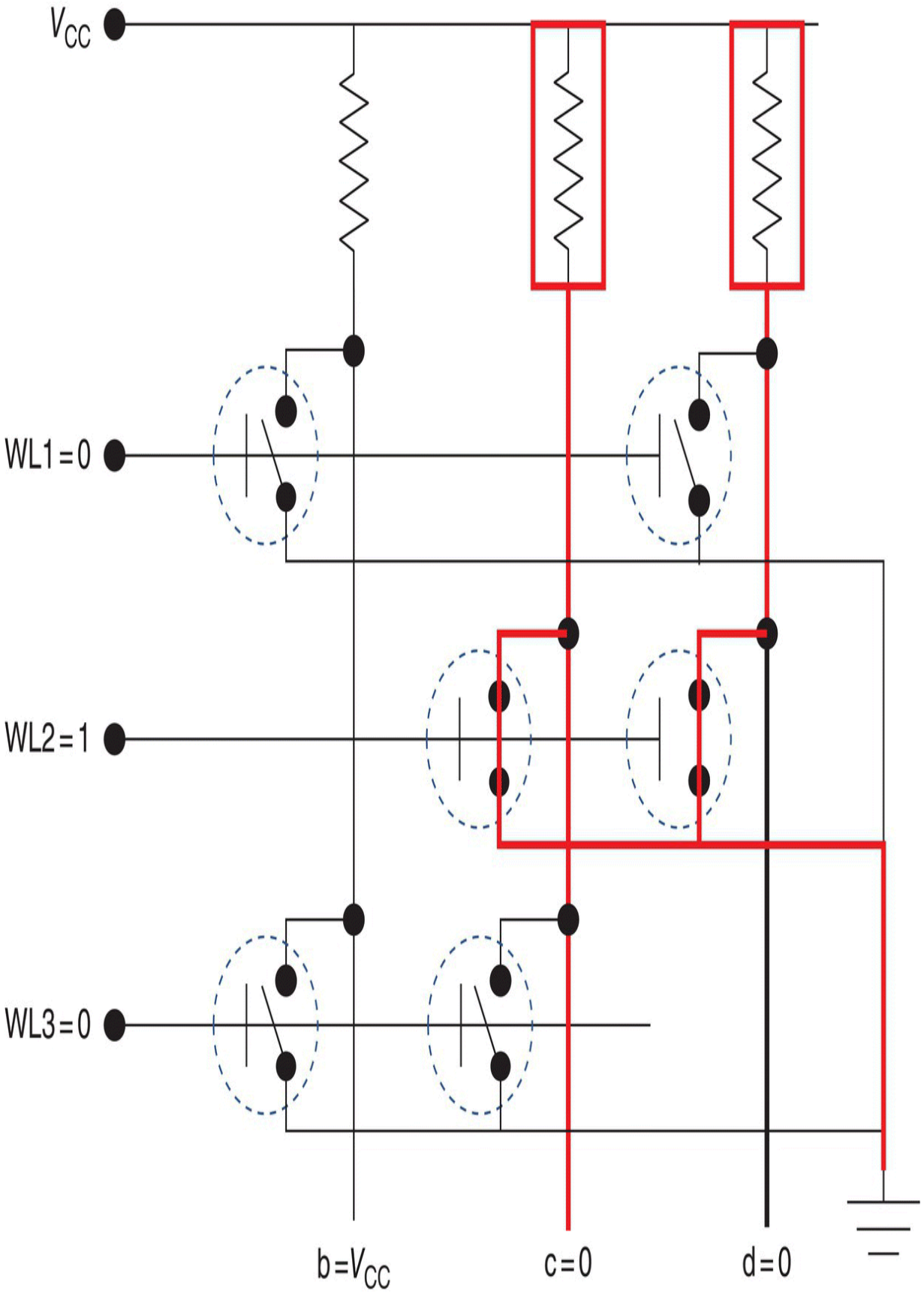

V_{CC}
WL1=0
WL2=1
WL3=0
b=V_{CC}
c=0
d=0

Figure 12.20 Switch representation of the ROM when one of the word lines, WL2, is ON. Only bit line b is not shorted to ground.

If we make the matrix larger and larger, we keep on alternating the location of the CMOS so that no bit line has the same gate connections to the word lines. A single CMOS in a bit that is shorted forces the bit line to ground, that is 0. That is also the reason why we have a resistor between the source voltage and the bit lines, so as not to short the b input voltage to ground. The CMOS positions are permanently fixed and cannot be changed. That is why this memory is read only.

12.5.4 Programable Read-only Memory

I show one type of programmable read-only memory (PROM) in Figure 12.21. In a PROM we start with an array of CMOS, which is much easier to design than a ROM, but instead of a wire connection between the drains and the bit lines we now have a fuse. If we apply a high current through the desired CMOS by selecting one bit line and one word line, we burn the fuse and thus disconnect the specific MOS. To replicate the ROM I show in Figure 12.19, we would select the second CMOS from the first word line and run a higher current through it to destroy the fuse and thus it appears as if the CMOS was never there. That is where the term “burn the ROM” comes from.

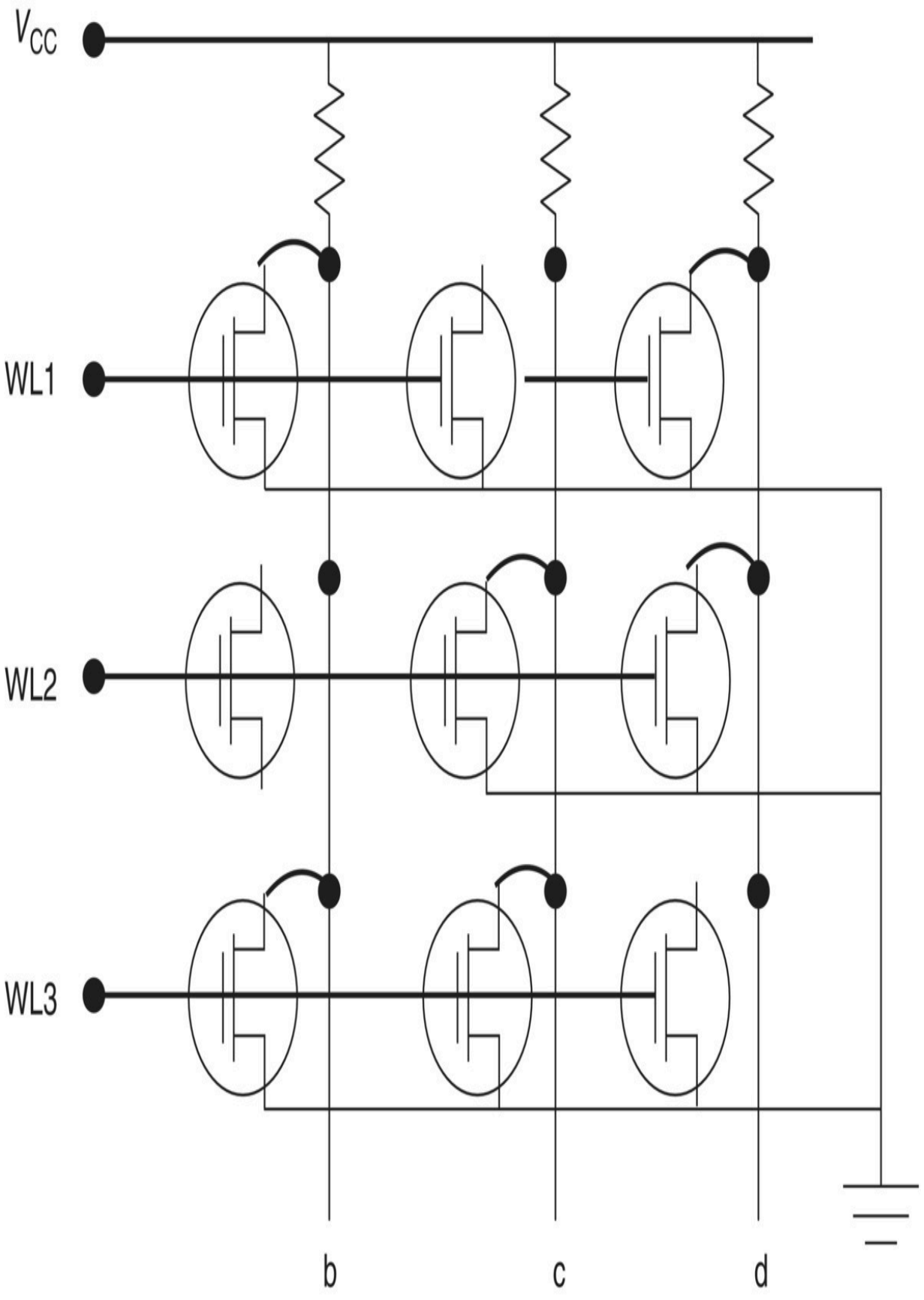

V_{CC}
WL1
WL2
WL3
b
c
d

Figure 12.21 A PROM has fuses connecting the sources to the bit lines. These fuses can be blown, leaving a permanently programed memory chip.

Another trick is the use of an extra gate between the gate that is connected to the word line and the CMOS channel (Figure 12.22). This extra gate is connected to nothing and it is just floating surrounded by insulating oxide layers. If we apply a large voltage between the gate that is connected to the word line and the source, electrical charges can "tunnel" through the thin oxide between the channel and the floating gate and charge the floating gate (one more use of the quantum mechanical theory of tunneling!). These floating gates can retain the charge for several years. When we apply a voltage at the gate, the intermediate charge of the floating gate stops the field lines and thus the CMOS is non-responding. These cells are called erasable programable ROMs (EPROMs). To erase them, we just apply the voltage in the opposite polarity and subject the chip to ultraviolet light, which ionizes the oxide and discharges the floating gates. The need to use ultraviolet light means that the whole memory matrix is completely erased, ready to be programed again. It is difficult to do this in a packaged ship.

Ready for another one? The electrically erasable programable ROM (EEPROM) does exactly what it says. It is the most versatile of all the memories. You do not need ultraviolet light. This is an advantage because you can erase the memory without removing it from the device to shine the ultraviolet light on it, and it can erase selected cells. The disadvantage is precisely that. It has to be done one cell at a time, so it is a slow process. To write we apply a high voltage at the gate and turn ON the selected bit line. This charges the floating gate. To read, we again select the bit line but now we decrease both the gate voltage and the word voltage so the selected bit line can read the state of the floating gate. To erase, we just turn OFF the bit lines and apply a large voltage at the source, and the charges in the floating gate tunnel back to the source, emptying the floating gate. Cell phones use EEPROMs to store system data so that providers can

reprogram the phone's program remodly. SIM (subscriber identity module) cards are also microprocessors with an EEPROM.

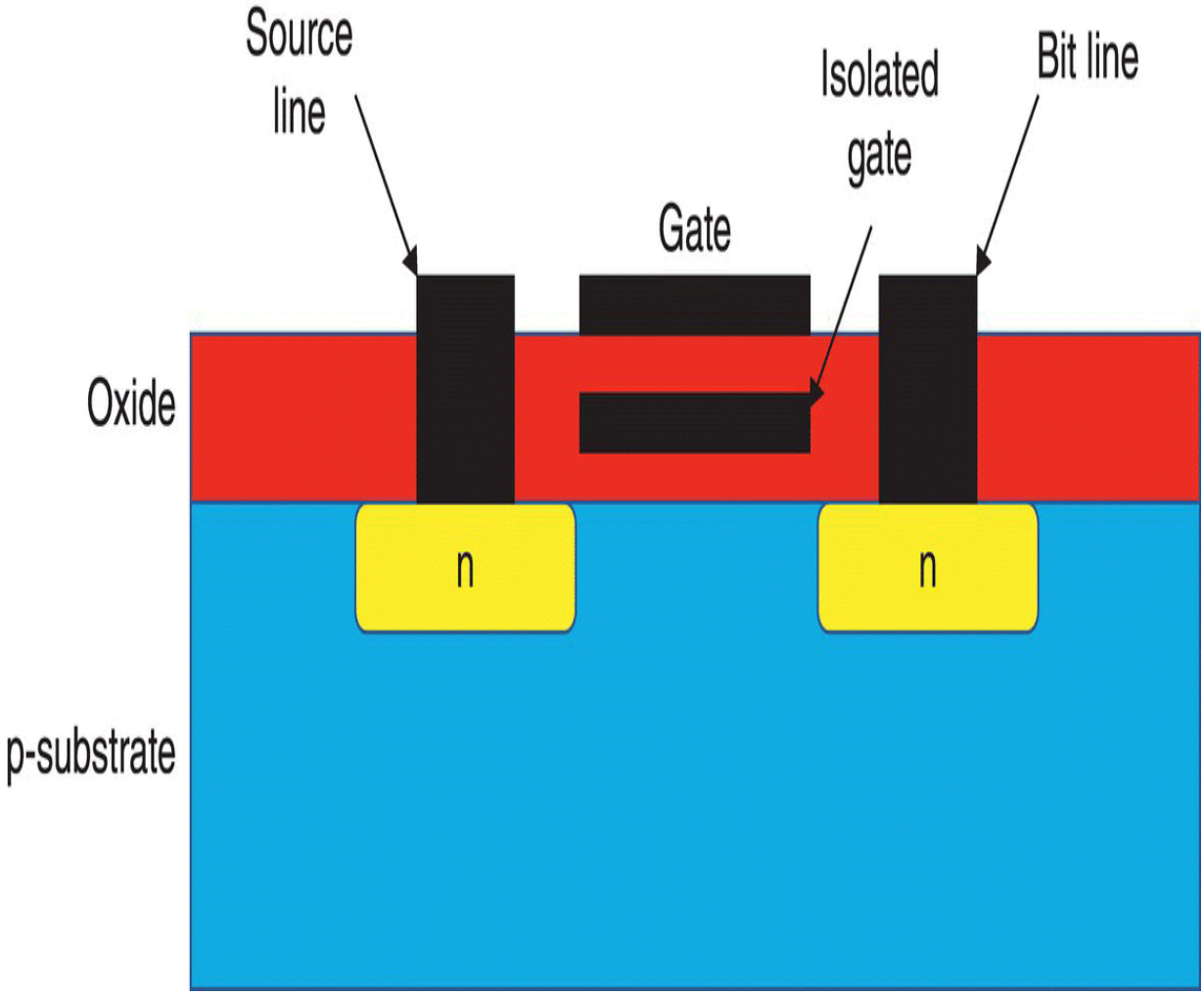


Figure 12.22 The EPROM consists of a regular MOSFET with a completely isolated gate between the electric gate and the substrate that can be charged by tunneling though the thin oxide between the gate and the channel.

12.6 Gate Arrays

There are many situations in which we need sophisticated chips but in relatively small numbers. This occurs in highly specialized programs such as many of the NASA programs, including the chips used in astronomical observatories, which require very complex and

specific chips to operate the electronics and the optical devices, such as infrared detectors. It would be awfully expensive if we had to run the whole process, mask included, just for the few chips we need.

Gate arrays solve this problem. They are the electronic designer solution, similar to the electrical kits we can buy at a toy store for our children. These are chips with multiple devices, logic modules, transistors, MOSFET, resistors, and capacitor, memories all strategically located in different parts of the chip. The only thing they do not have is many of the connections. The user buys a few wafers, the gate arrays. He designs the aluminum or polysilicon masks to make the connections, and only the last few masking steps are needed to complete the design. These gate arrays come with computer programs that help you to design the interconnects. This makes a process that would cost millions of dollars into a more affordable (still in the hundreds of thousands of dollars range) process. It is also a much faster process than if we were to start from scratch.

12.7 Summary and Conclusions

In this chapter I explain many of the large components that are used in all computers and microprocessors, and practically all electronic control boards. These are the multiplexer and all types of memory for different uses, some random access (RAM) and some not (ROM), some programable (PROM) and some not, some faster (SRAMS) and some slower (DRAMS). In Chapter 14 I take all of these components and talk about the computer and the microprocessor.

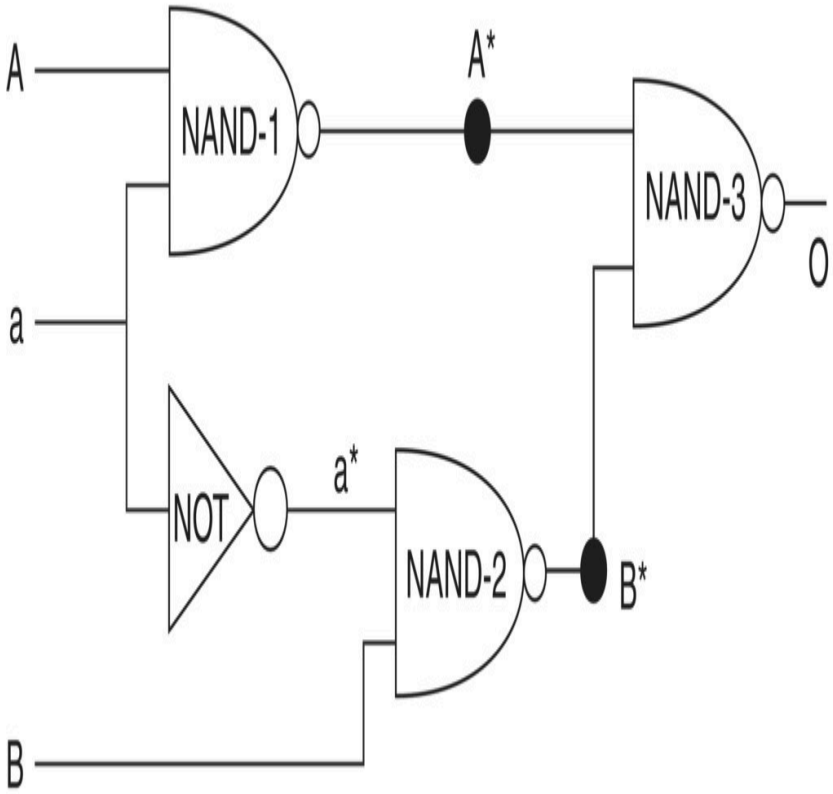
Next, I digress (again!) and discuss another very important use of semiconductors, optoelectronics, the interaction of light and semiconductors.

Appendix 12.1 A NAND implementation of a 2 to 1 MUX

I mentioned several times when explaining the circuit implementation of a given function that there are several different ways to obtain the same result. As an example, here I discuss a different implementation of the MUX (Figure 12.23). This is actually a preferred implementation because it is easier to lay out and it is a little faster. Let's see how it works.

Notice that all the elements are NANDs, the negative of AND. Also, the NOT module ensures that when the control line a is 0 only NAND-1 is able and NAND-2 is not, and vice versa. When the control line a is 0, the output is equal to A and when a is 1 then the output is equal to B.

Let's look at it more carefully. First, remember that the output of a NAND is 1 only when both inputs are 0. So when the control line a is 0, a* is 1 and therefore the output of NAND-2, B*, will be 0 no matter what the value of the input B is. If A is 0, A* will be 1 and the output of NAND-3 will be 0, the same as A. When A is 1, A* is 0 and the output O is back to 1. That is, when the control line a is 0, the output is exactly equal to whatever the value of A is. You can easily see that when a is 1, NAND-1 is disabled and the output is equal to the value of B.



a	A	B	O
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

Figure 12.23 Implementation of a 2 to 1 MUX using three NANDs and one NOT module.

13
Optoelectronics

OBJECTIVES OF THIS CHAPTER

Although I have already discussed some of the optical electronic devices when I talked about infrared detectors in Chapter 4, I'd like in this chapter to bring together semiconductor devices that interact with light, topics that include photoconductors, lasers, and light-emitting diodes. I cover these devices in several sections, which all involve the effect of light on semiconductor and pn-devices.

13.1 Photoconductors

The simplest optoelectronic device is the photoconductor. It consists of a piece of semiconductor material and just two contacts (Figure 13.1). Intrinsic semiconductors have very few electrons in the conduction band. In Section 2.4 I mentioned that intrinsic silicon has 1.45×10^{10} electrons per cm^3 in the conduction band, which is a tiny number compared to the total number of silicon atoms, 5×10^{22} atoms per cm^3. That means that the resistivity of pure silicon is 60 000 Ω-cm. If we apply a voltage and light shines on the semiconductor, many electrons absorb the light and generate free electrons (and holes), the resistivity of the semiconductor decreases, and the ammeter measures the change in current. Since I have spent considerable time explaining infrared detectors in Chapter 4, I do not need say more here. The semiconductor lets me know that light is shining by the change in current: the more current, the stronger the light.

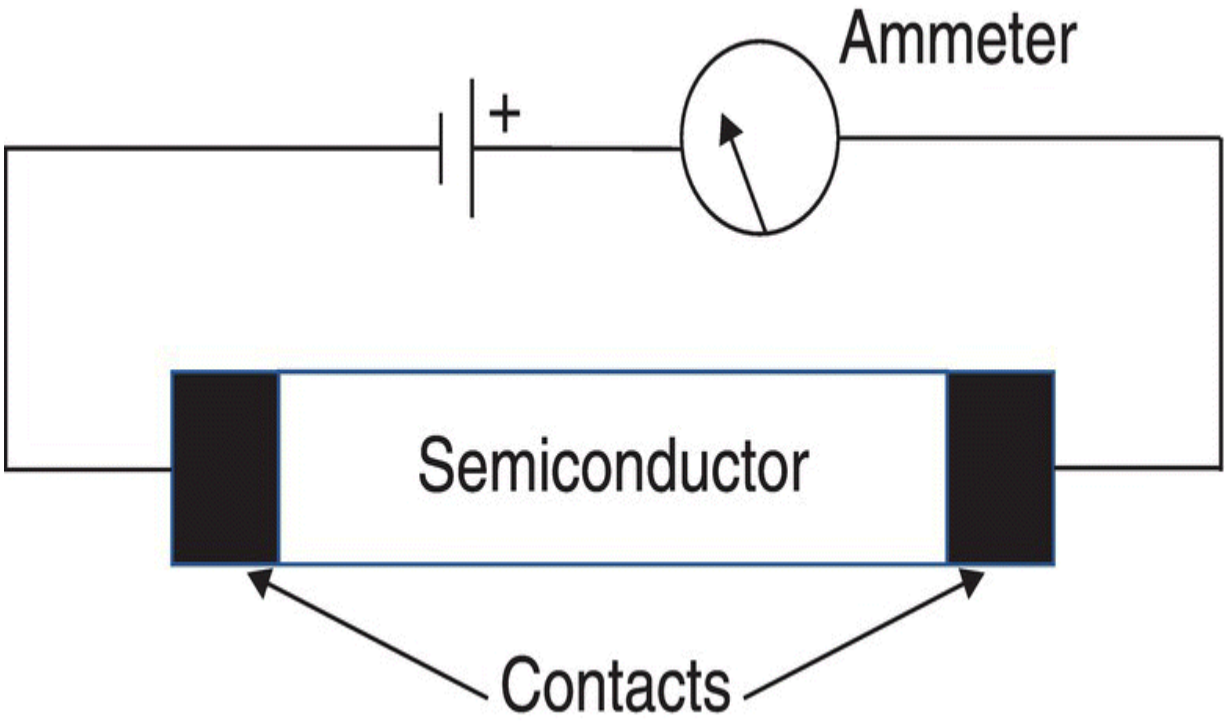


Figure 13.1 A simple photoconductor consists of a semiconductor with two contacts.

13.2 PIN Diodes

The PIN diode is a pn-junction with a large intrinsic region between the p- and n-semiconductors. Figure 13.2 shows what happens when a reversed biased diode is illuminated with light with an energy higher than the energy gap of the semiconductor, *Eg*. The semiconductor absorbs the light energy and kicks an electron from the valence band up to the conduction band as long as the energy of the radiation is larger than the energy gap, *Eg*. If the photon is absorbed in the two bulk p- and n-type regions of the semiconductor, cases 1 and 5, in Figure 13.2 the electrons and holes move slowly in the region but eventually recombine. They do not contribute to any added current. In case 3, when the photon is absorbed in the transition region and creates an electron–hole pair, the internal electric field moves the electron to the positive side and the hole to the negative terminal. We have now created an excess charge that is measurable and thus contributes to a signal. Cases 2

and 4, where the electron pair is created close to the transition region, can work both ways: some of the electrons and holes are absorbed by the bulk semiconductor but some may drift to the transition region and be affected by the electric field and thus contribute to the external current.

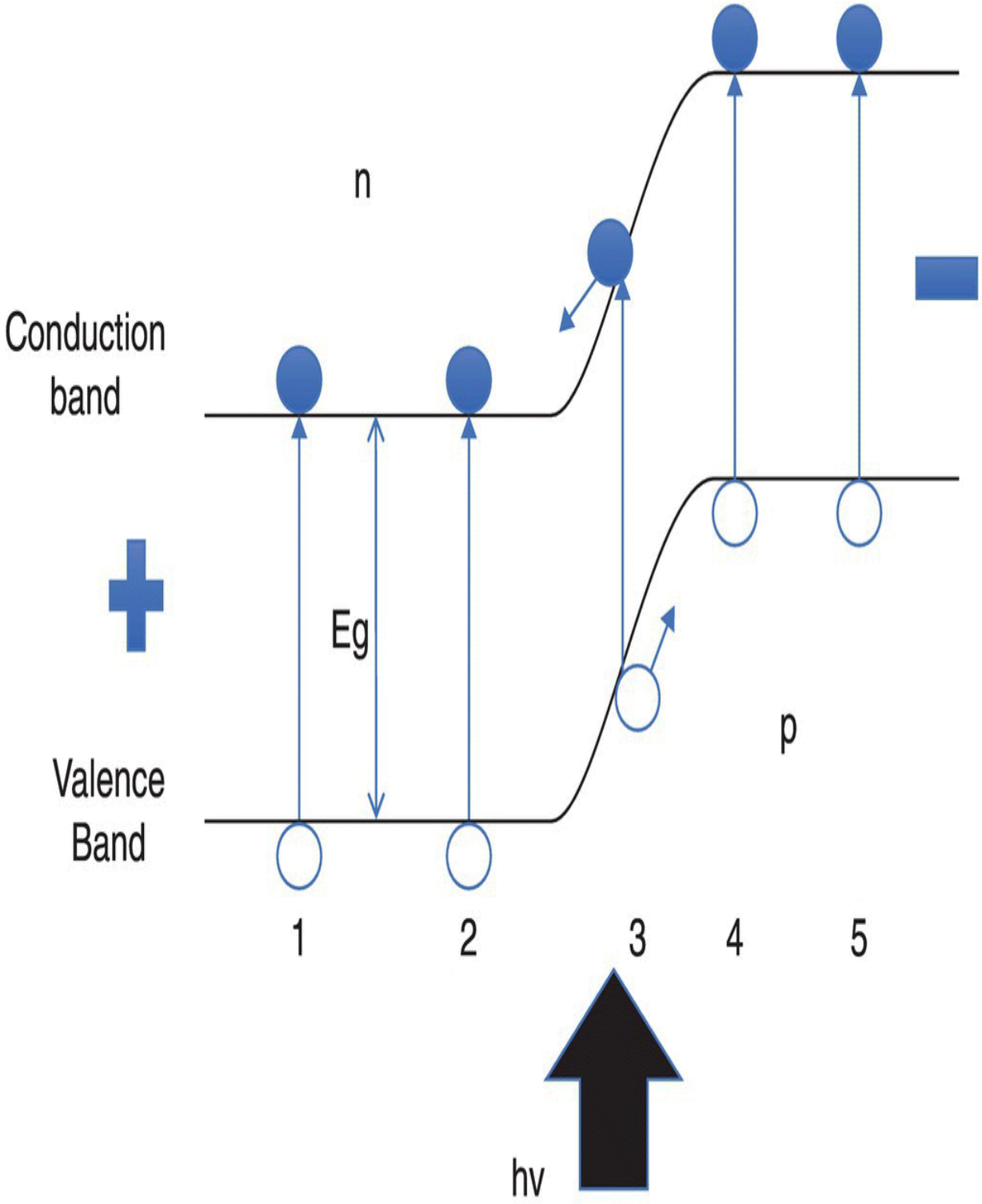


Figure 13.2 Radiation shining on a reversed-biased diode creates an electron–hole pair and the charges are swept in opposite directions by the internal electric field.

One problem with this structure is that the transition region is very narrow compared with the entire semiconductor diode. Thus, only a very small number of the photons created by the light affect the change of the current. That is where the PIN diode comes to the rescue (Figure 13.3).

We still have the pn-junctions, but the p and n regions are separated by a large intrinsic region, so when we reverse bias the junctions there is a potential created by the intrinsic region that acts like a resistor. Any electron–hole pairs generated by the photons that are absorbed in the intrinsic regions as well as in the two transition regions are separated, the electrons moving toward the positive terminal and the holes toward the negative one. What we have accomplish is to make the region where the internal electric field separates the charges much longer. Now the active region is not restricted to the transition region but has expanded from A to B.

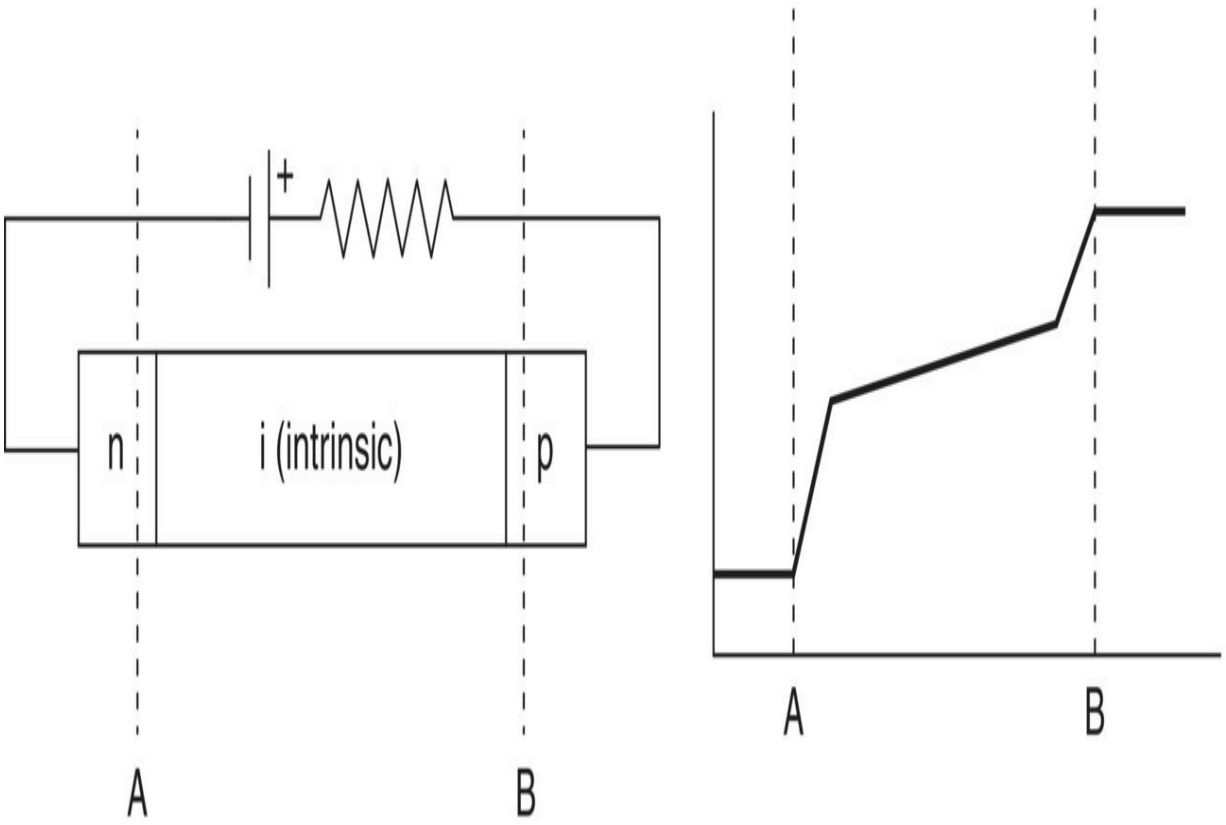


Figure 13.3 The PIN diode structure consist of a p- and an n-region separated by a large intrinsic region (left) creating a much longer electric field region (right).

13.3 LASERs

13.3.1 Laser Action

In 1953 Charles H. Townes (1915–2015), a professor at Columbia University, invented the first MASER, which stands for microwave amplification by stimulated emission of radiation, a mouthful but very descriptive. (The MASER was an experimental and esoteric device when Townes first proposed it and the cynics said that MASER stood for money acquisition scheme for experimental research. When the LASER became popular, the same cynics call it the large acquisition scheme for experimental research. This shows that many physical concepts and experiments seem, at the beginning, to be of no use whatsoever, just a toy for the scientist to

entertain themselves and get research grants. What if funding had not been available for this highly experimental research?) LASER stands for light amplification by stimulated emission of radiation. The operation of both MASERs and LASERs uses the energy level concepts of the Bohr atom (Figure 13.4). First, we pump electrons from the ground energy level to an excited level. The excited state is not stable and its retention time is very short, in the microsecond range. Very quickly the electrons fall to a very stable level (called a metastable level). They stay at this metastable level for a while (milliseconds), creating a population inversion, that is, many more exited electrons with energy larger than the ground level. Then one of the electrons in the inversion layer spontaneously comes down to ground level, releasing its energy in the form of a photon. This generates a wave with a frequency of energy equal to the difference between the ground level and the inversion level. This wave is inside a cavity with reflecting ends and as the wave goes back and forth inside the cavity it stimulates other electrons from the inversion layer to go down to the ground level with exactly the same frequency and phase as the first electron, thus creating a larger and larger beam in which all the rays have exactly the same wavelength and phase. This is what Townes did with ammonia molecules, which spontaneously emitted waves in the 24 GHz frequency, the microwave range.

Coherent light means that the light is monochromatic, that is, all the rays have exactly the same frequency and the same phase so the peaks and valleys of all the waves occur at exactly the same time. At the left of Figure 13.5 I show three waves of a coherent light; all have the same frequency and the same phase. They fit perfectly one on top of the other. The waves of an incoherent light, on the right, have either a different frequency, compare wave A to wave B, or a different phase, compare A to C, or both, compare B to C. All the sources of light are incoherent. The light from a lamp or the sun consists of waves of different frequencies, but light generated by the difference between atomic levels is coherent.

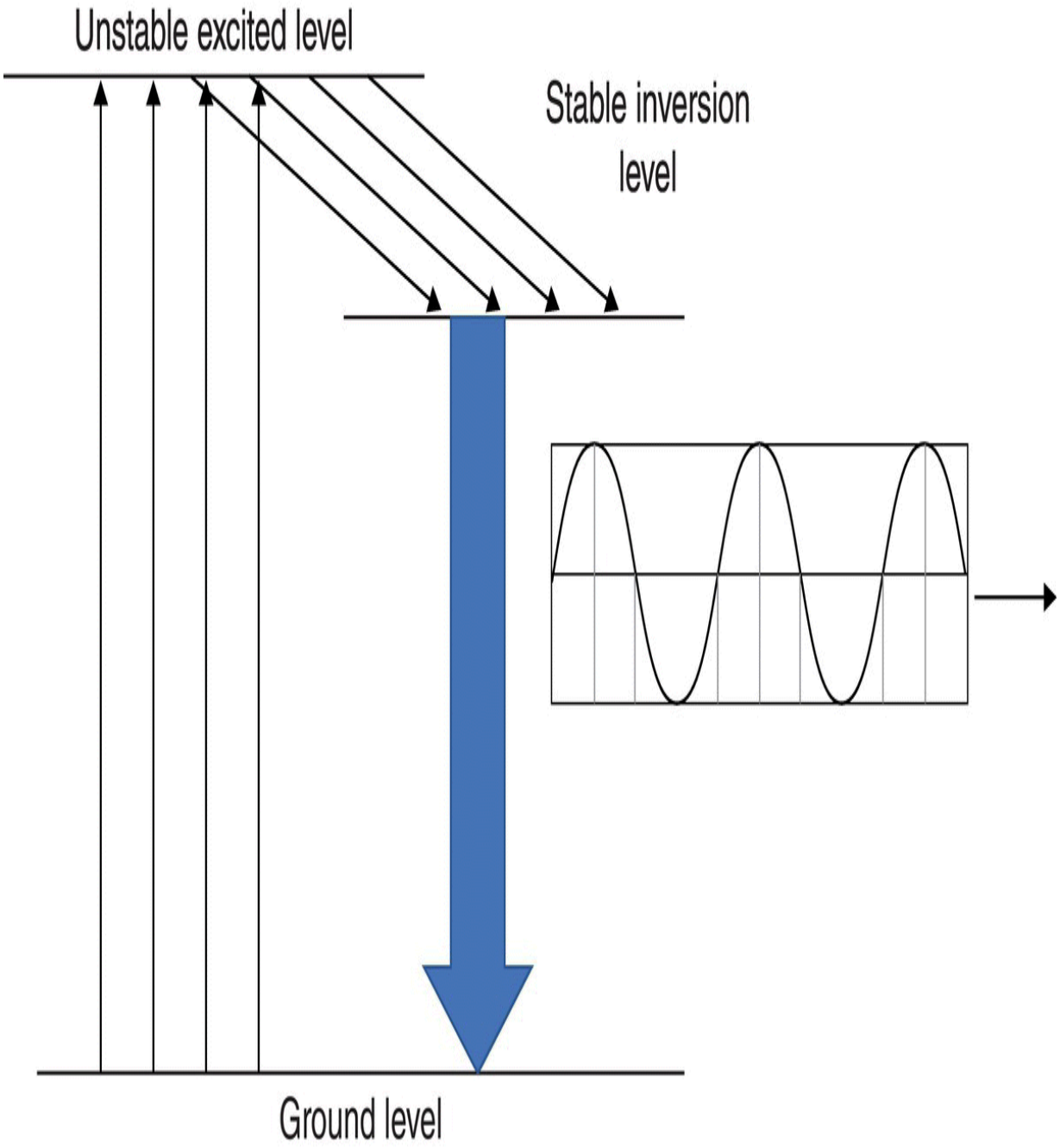


Figure 13.4 Both MASERs and LASERs work with the idea that electrons that are excited to an unstable level decay quickly to a stable one where they accumulate in an inversion level. When triggered, the electrons fall to ground level, generating a coherent light.

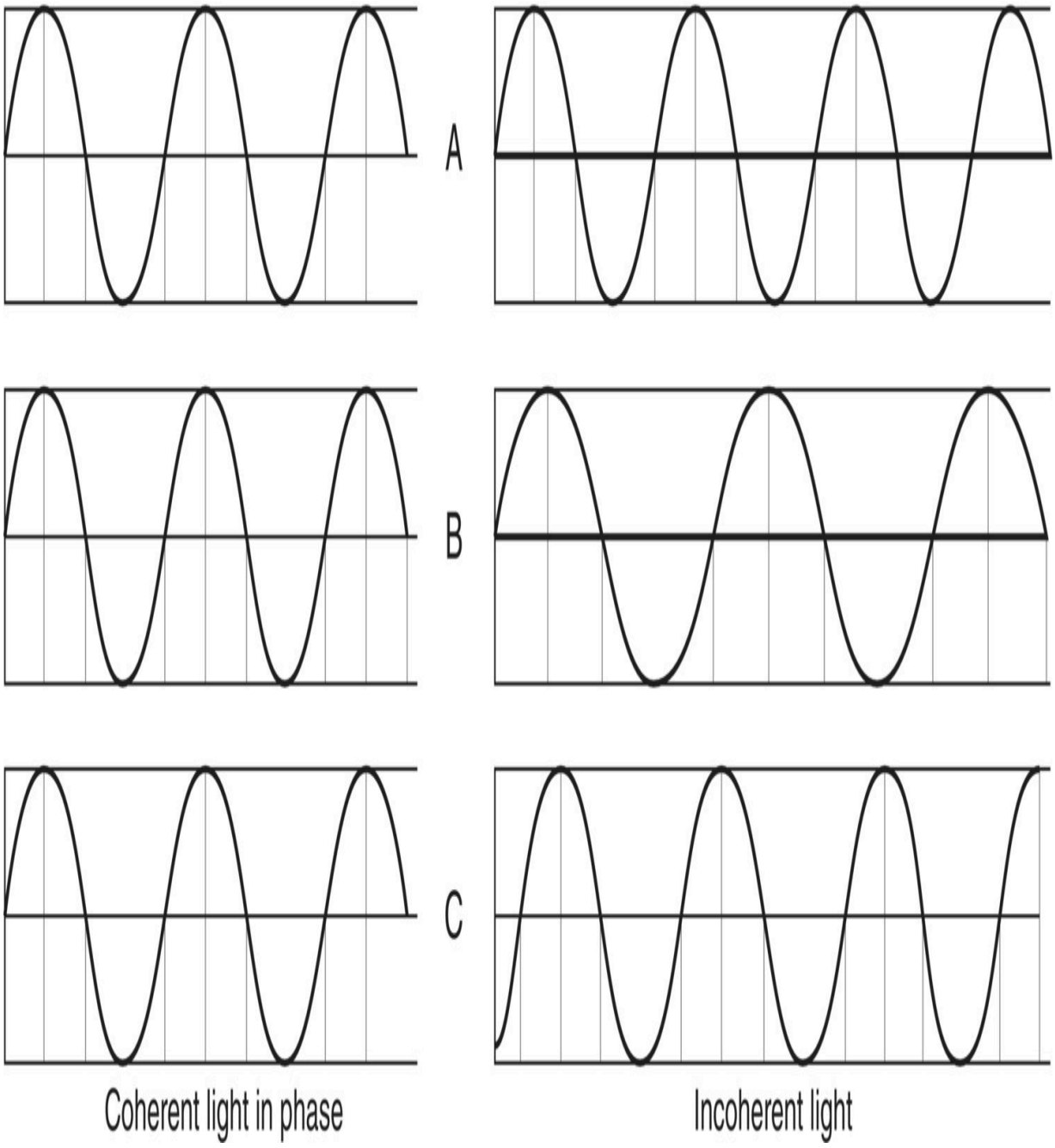


Figure 13.5 In a coherent light (left) all the waves A, B, and C are exactly the same. In an incoherent light (right) the waves have either a different frequency (A and B) or a difference phase (A and C) or both (B and C).

The first thing we need to operate a LASER is to create a population inversion. This does not happen naturally. The electrons will always want to be at the lower energy levels. So, we need something to

excite the electrons. The terminology is “energy pumping.” Bright pulses of light from flash tubes do the trick.

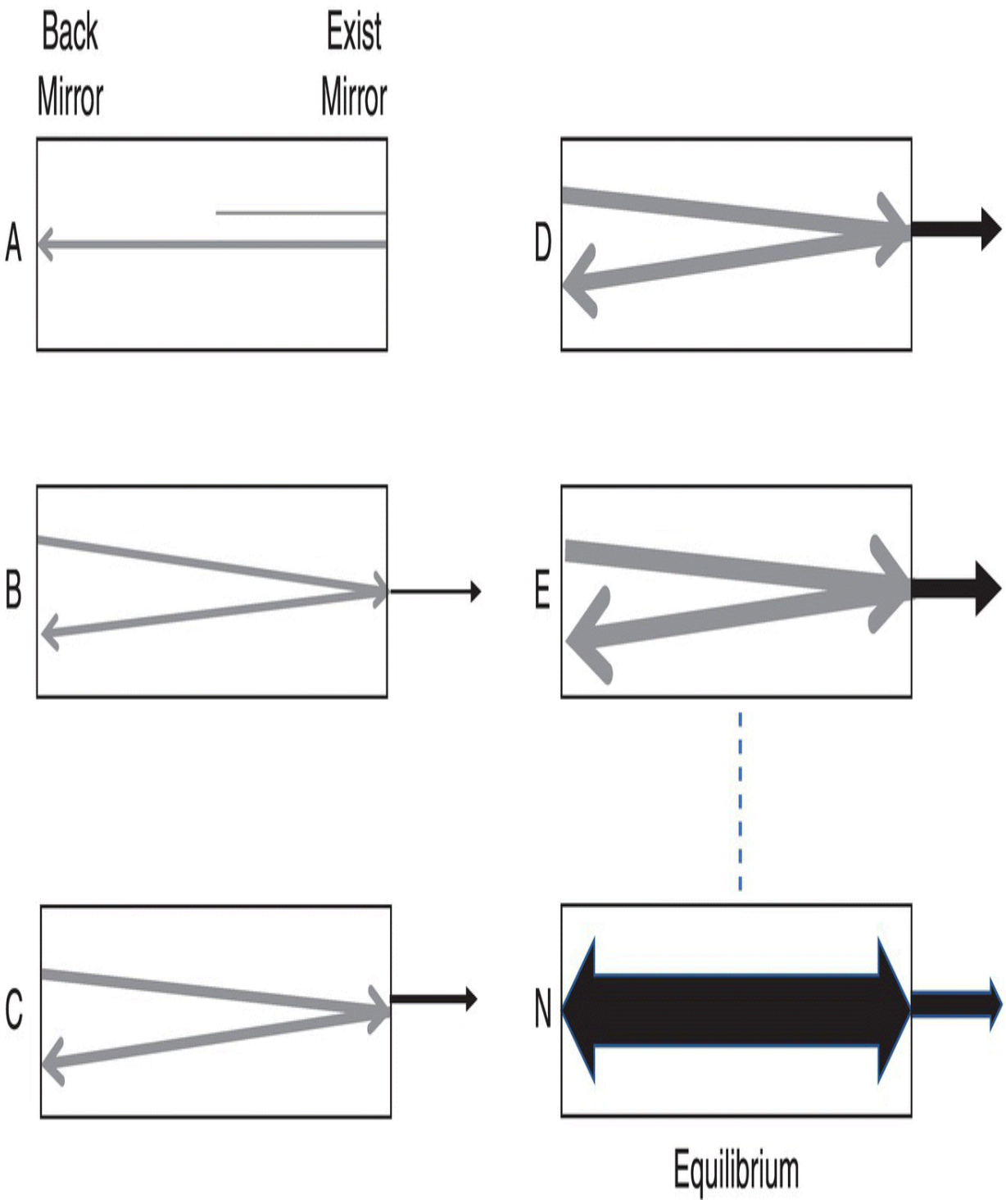


Figure 13.6 The beam of light bounces inside the cavity with one fully reflective mirror on the left and a partially reflective mirror on the right. Only a small portion of the light escapes the cavity.

The next thing we need for a LASER to work is a cavity where the

electrons can move back and forth, bouncing off the end walls, and thus trigger more electrons to fall down to the ground level at exactly the right time. The cavity has to have the right length so that the wave resonates going back and forth, exciting additional electrons to drop and donate their energy to the rays. Mirrors at both ends generate the resonant cavity. A ideal resonant cavity must fulfill the relationship

$$n\lambda = 2L \qquad (13.1)$$

where n is an integer number, λ is the wavelength of the LASER-generated photons, and L is the length of the cavity This ensures that the waves will go back and forth, adding up and thus amplifying the beam. Suppose now that one mirror is totally reflective, but the second mirror is only partially reflective, let's call it the exit mirror. The exit mirror may reflect from 20% to 95% of the light, thus transmitting 80% or 5% of the light.

When we flash the light, many electrons move to the exited level and very quickly fall to the inversion level. Now either a trigger from the outside or a spontaneous decay initiates the beam of the exact frequency demanded by the location of the atomic levels (see Figure 13.6). As the process starts (case A), very few electrons go back and forth, but each time they travel through the cavity they trigger other electrons to jump back down and the energy lost by the electrons is added to the beam (case B). Now 5% of the beam escapes the cavity through the partially reflective mirror and 95% goes back. This continues (cases C, D, and E) with both the internal beam and the exit beam growing until they reach an equilibrium condition, N. The flow of the coherent light out of the cavity continues as long as we keep exciting electrons. In pulse LASERs, the beam stops as soon as all the electrons have come down to the ground level.

To work perfectly, the length of the cavity should be equal to an exact number of wavelengths. The problem is that the cavity is large, maybe a foot long, and there are probably 500 000 wavelengths inside the cavity. How can we possibly satisfy exactly

the condition of Eq. (13.1)? Of course, we can't. So, in addition to the primary mode there other very close modes that are resonant, but the magnitude of these secondary modes decreases in amplitude quickly and only the main one is really strong. At the exit we can use filters and optics to remove the unwanted rays.

13.3.2 Solid-state Lasers

Besides the gas LASERs I discussed in the previous section, we also have solid-state LASERs. The lasing materials are solid materials with some impurities that are the lasing atoms. For example, the ruby LASER contains chromium atoms indented in sapphire material and it is the chromium that is lasing (the chromium atoms also give the ruby its red color). The operation of these solid LASERs is very much like that of the gas LASER (Figure 13.7). A flash lamp surrounds the ruby rod and excites electrons. As in the gas LASER, some of the electrons spontaneously decay and generate red photons. These photons move back and forth inside a cavity and trigger more electrons to decay to the ground level, generating more photons in the process. There are very few chromium atoms inside the ruby and therefore they act like the impurity atoms in semiconductors.

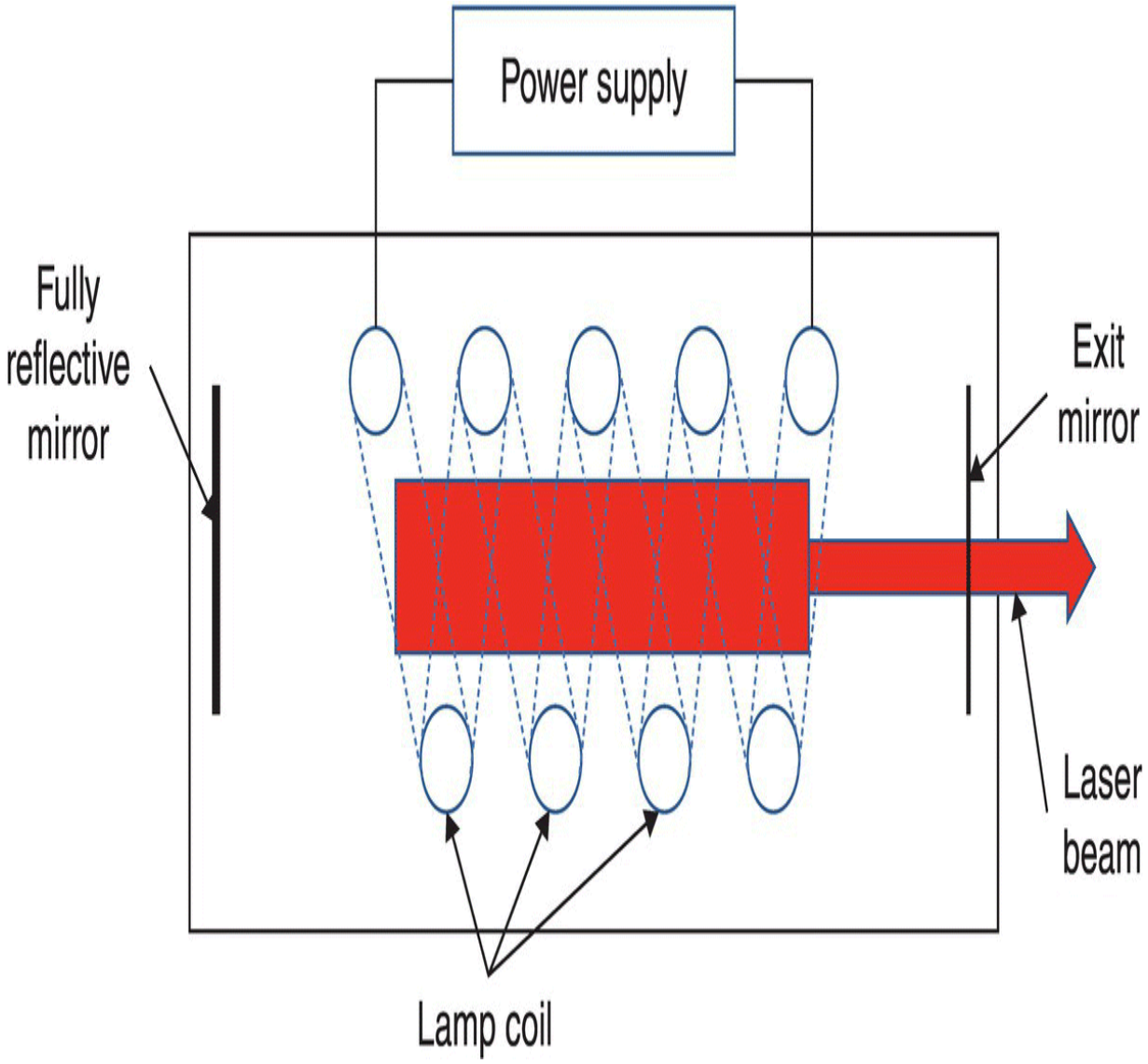


Figure 13.7 A ruby LASER in a reflective cavity surrounded by a light coil that provides the energy to create the population inversion.

13.3.3 Semiconductor LASERs

This book's objective is to understand how and why semiconductors work, so now that I have used gas and ruby LASERs to explain how LASERs work, let's talk about semiconductor LASERs. We have seen that to have a LASER we need three things: a population inversion, gain, and a resonant cavity.

In Section 3.4, I mentioned that for normal diodes and transistor devices we want to have very few selected impurities so that they act like isolated atoms inside the silicon so as not to start interacting with each other creating an energy band of their own. Toward the end of the same section I mentioned that in some cases we want to have a very large number of impurity atoms (e.g. when we need to make good contacts to aluminum lines, as we saw in Section 10.7.1). A laser is a device that uses degenerate semiconductors. Figure 13.8 shows the same figure as Figure 5.3 except that because I have a large number of impurities in both the n- and the p-type semiconductors, the transition region is steeper and narrower, and the voltage generated internally by electrons and holes moving across the transition region is much larger. (We saw something similar with the tunnel diode in Figure 5.14. I explain why the transition region thickness is thinner when the concertation of impurities is larger in Appendix 5.3.)

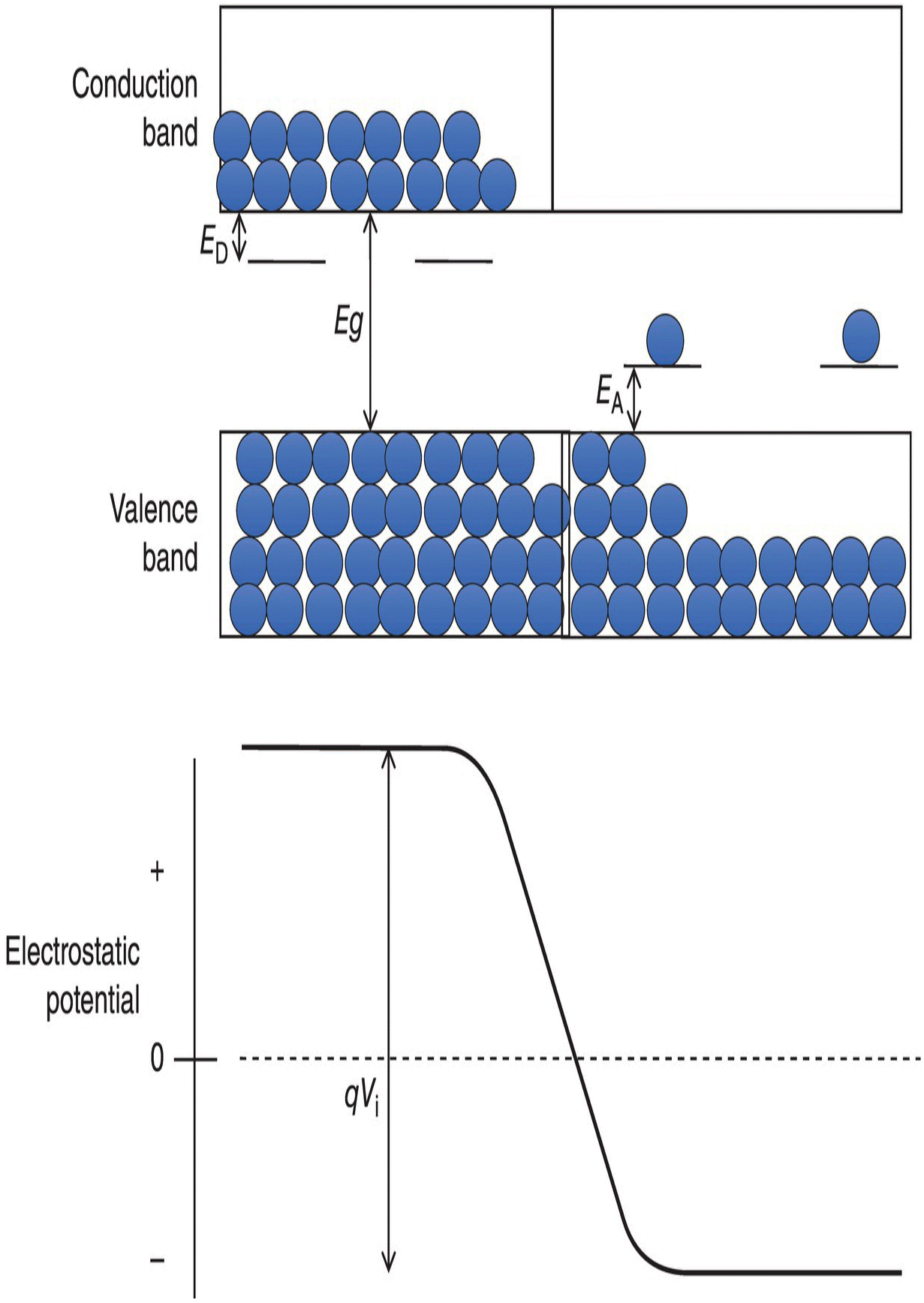

Conduction band
E_D
Eg
E_A
Valence band
+
Electrostatic potential
0
qV_i
–

Figure 13.8 The internal voltage for a degenerate semiconductor diode is large and the transition region is quite narrow.

Figure 13.9 shows what happens when we forward bias this p-n junction. Take a look first at the sketch at the left. The n+ degenerate semiconductor has lots of electrons in the conduction band (I show them in solid black) because we have a lot of n-type impurities. I show the p+ regions as a white region at the top of the valence band. Again, because I have a lot of p-type impurities, I have created a lot of holes at room temperature. The transition region is quite narrow and the slope of the electrical potential is quite steep. What happens if I forward biased this junction, as I show on the right of Figure 13.9? The voltage between the p+ and n+ region decreases. At the edge of the transition region, on top of the holes in the valence band of the p+ side, there is a large number of electrons from the conduction band of the n+ side. We have created an inversion layer, with a large number of electrons in a higher level with the desire to go to an allowed lower energy level. When the electrons start coming down, they emit a wave with energy equal to *Eg*.

The second thing we need is a resonant cavity. Notice that in the p-n junctions the transition region is tiny compared with the bulk semiconductors that create the junction (Figure 13.10). The generation of photons occurs in the transition region, as I show in the sketch on the left and the output flows out from the sides. You can say, yes, the photons are created in the transition region, but why don't they go in all directions? They do, but two things happen. The sketch on the right shows an expanded view of the transition region. The small circles in the transition region represent photons created by an electron falling to the valence band of the p+ semiconductor. Highly doped semiconductors have a small difference in the index of refraction, therefore some of the photons that are directed toward the semiconductor are reflected back into the narrow transition region. Photons 1, 3, 4, and 5 have a small angle, and are reflected back into the transition region. Photons 2 and 6 reach the edge of the transition region with an angle greater than

85° and they are refracted to the semiconductor bulk. They are absorbed almost immediately in the bulk of the semiconductor and they neither contribute to nor mix with the photons that move horizontally inside the transition region. Therefore, not only is the beam coherent, all electron–hole pair combinations release a photon of exactly the same frequency, but it is also collimated with a very small dispersion angle.

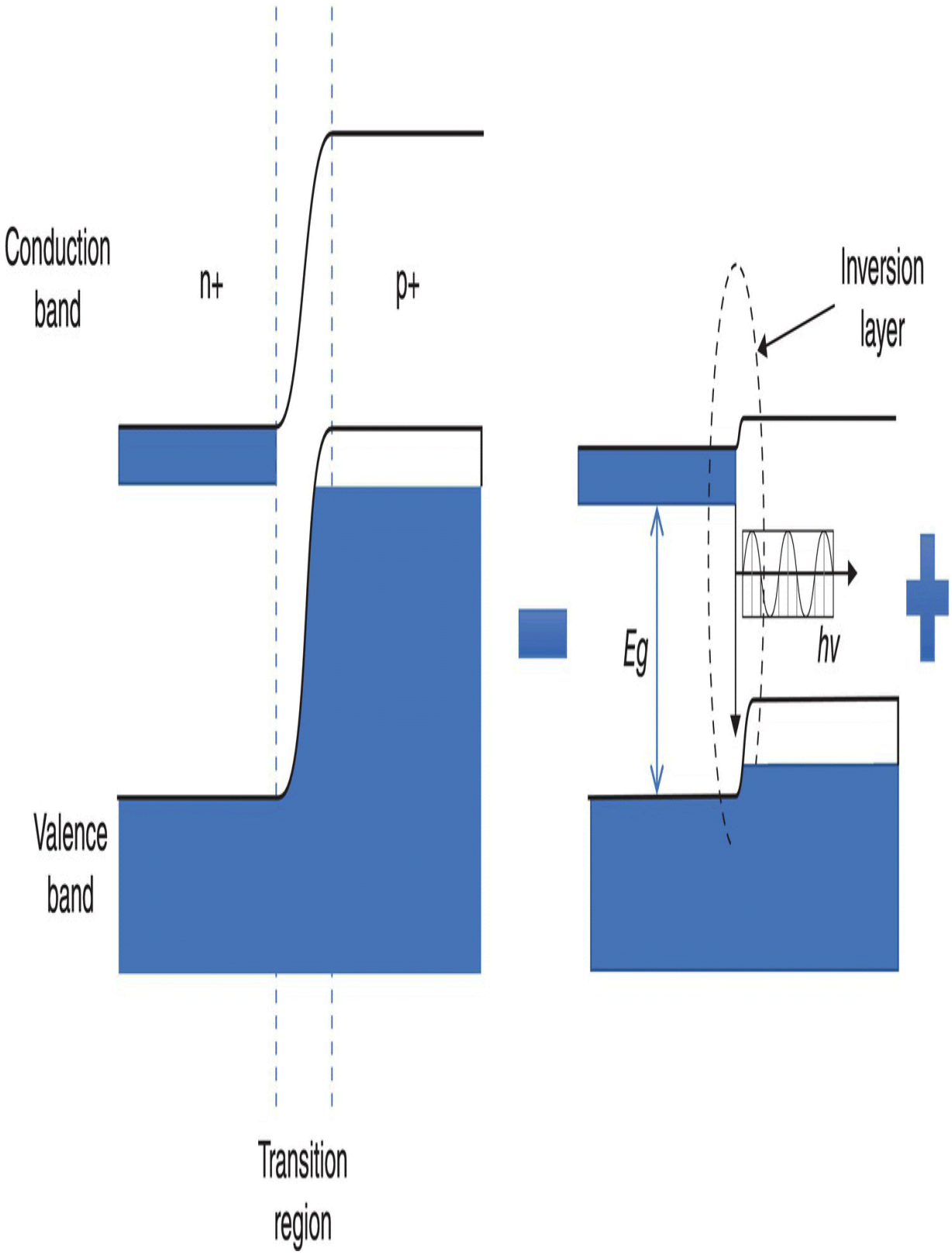

Conduction band
n+
p+
Valence band
Transition region
Inversion layer
Eg
hv

Figure 13.9 On the left we have a highly doped pn-junction. When we forward bias the pn-junction (right) we create an inversion layer.

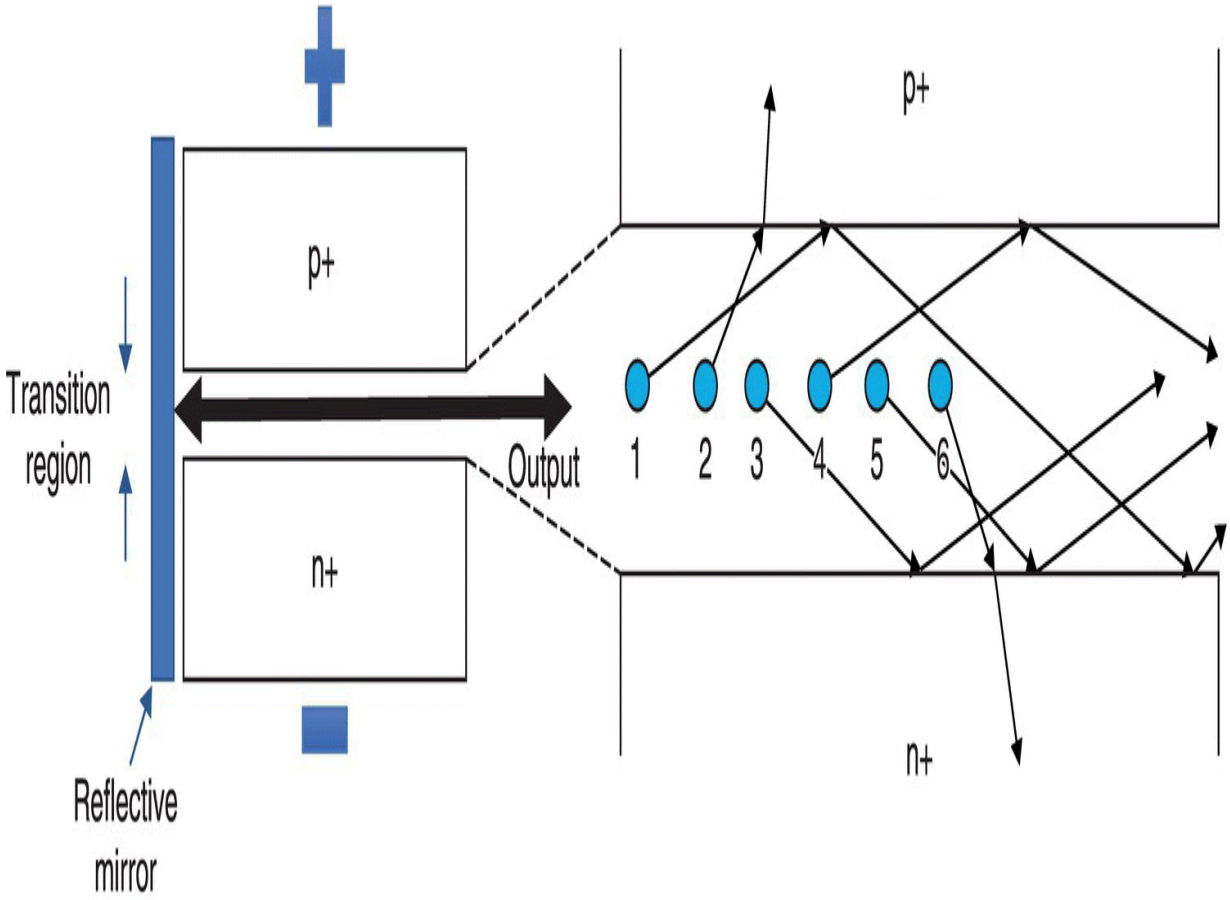


Figure 13.10 In a LASER semiconductor, the reflective properties of the transition region itself form the required LASER cavity.

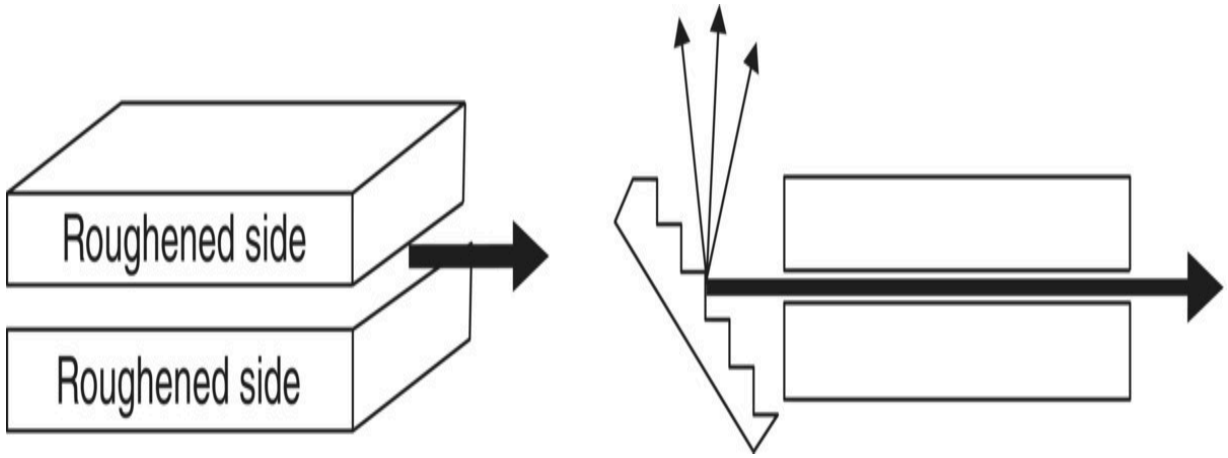


Figure 13.11 Some methods to confine the beam inside the semiconductor cavity and reflect the LASER beam back and forth within the junction.

The final condition for lasing is gain. This is easy to explain. The photons as they move through the transition region excite more and more electrons to drop from the conduction into the valence band and generate additional photons.

The three LASER conditions have been fulfilled: population inversion, gain, and cavity.

There are several tricks to confine the beam inside the transition region. The refractive index of silicon is 3.98 and that of GaAs (another common semiconductor material used for LASERs) is 3.6. This means that the flat, polished faces at the ends of the diodes reflect about 30% of the photons that impinge on them. This may seem a small amount but the high gain of the semiconductor LASER is sufficient to increase the strength of the LASER beam. The sides of the semiconductor can be roughened so that not too much light leaks through the sides (left-hand side of Figure 13.11).

The sketch on the right of Figure 13.11 shows a way to "tune" the semiconductor LASER. We apply a grading so that all frequencies not perfectly tuned are reflected away from the transition region and only the desired frequency is reflected back into the LASER cavity.

GaAs is the preferred material for semiconductor LASERs. Modern LASERs have complex junctions with four and five different layers of GaAs and AlGaAs, both p- and n-type, creating heterojunctions that are much more efficient to confine the LASER beam.

13.3.4 LASER Applications

LASERs have many applications. They can produce a continuous or pulsed beam of very concentrated light that can be focused very tightly on small areas. Some of the most common applications of LASERs are the following:

The most common application of LASERs is for writing and reading CDs, DVDs, and Blue-Ray discs (Figure 13.12). The LASER light is reflected by a beam splitter that reflects light that comes at an angle. The light is concentrated into the object to be scanned and it is reflected back through the beam splitter and into the light detector. The signal from the detector is sent to a processor that interprets the result.

By far the most familiar LASER application is in barcode scanners. A moving mirror scans the bar code and the reflected light goes back to a detector that interprets the bars and converts them into 1s and 0s, thus identifying the product. Barcodes can be linear or two-dimensional.

In LASER printers the LASER light discharges the static electricity of a drum. The drum collects ink from the toner only in the areas where there is static electricity and transfers it to the page. The number of dots per inch in modern printers is as high as 4000, which makes the copy as good as the original.

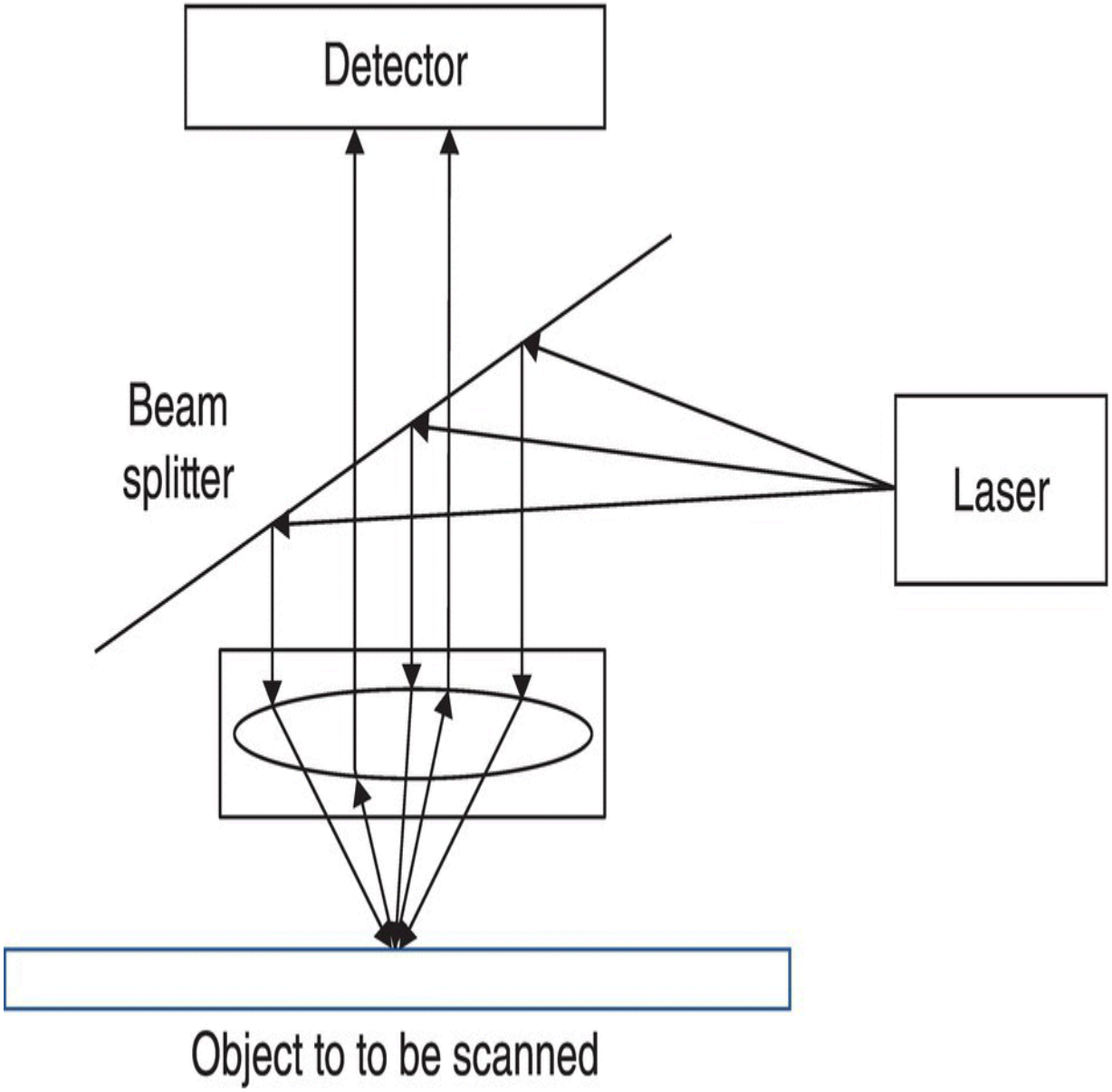


Figure 13.12 Typical system for LASER scanning.

LASERs are also the source of communication through fiber optics cables. The LASER light is modulated at the source, travels through the fiber optics, and is read at the other end, traveling at the speed of light.

LASER light is also used for drawing straight lines between walls and floors/ceilings so that construction workers can check that the walls or whatever else they are working on are perfectly level.

Pulse lasers are used to measure distance. The pulse bounces off the target and the range finder calculates the distance by looking at how long the pulse has taken to come back.

LASER pointers point to a figure or a word using a red dot. They can be attached to rifles, for example, to help the shooter to aim.

In industry high-power LASERs are used to drill and cut anything from aluminum to rubber. They can also weld two metal pieces together.

In Chapter 10 we studied how to fabricate integrated circuits. As the semiconductor processing minimum design rules get smaller and smaller, we use LASERs to both fabricate the photolithographic masks and illuminate the localized portions of the photoresist we want to remove.

In medicine doctors use LASERs in surgery. The LASER vaporizes cells without destroying any underlying tissues that you want to preserve, such as in eye cataract and cornea repairs. Dermatologists use LASERs for all types of skin conditions. In dentistry LASERs can replace the feared dentist's drill.

High-power LASERs can trigger nuclear fusion and the military uses high-power LASER in many applications.

Finally LASERs can be used to provide energy and kick electrons to the excited states of other more powerful LASERs.

13.4 Light-emitting Diodes

LEDs are very similar to LASERs in theory and operation. The main difference is that LEDs work based on spontaneous radiation, while LASERs use stimulated radiation, that is, LEDs just let the photons shine while lasers bounce the photons back and forth on a cavity to increase power and coherence. The result of the difference in operation is that LASER light is coherent and focused and LED light is not.

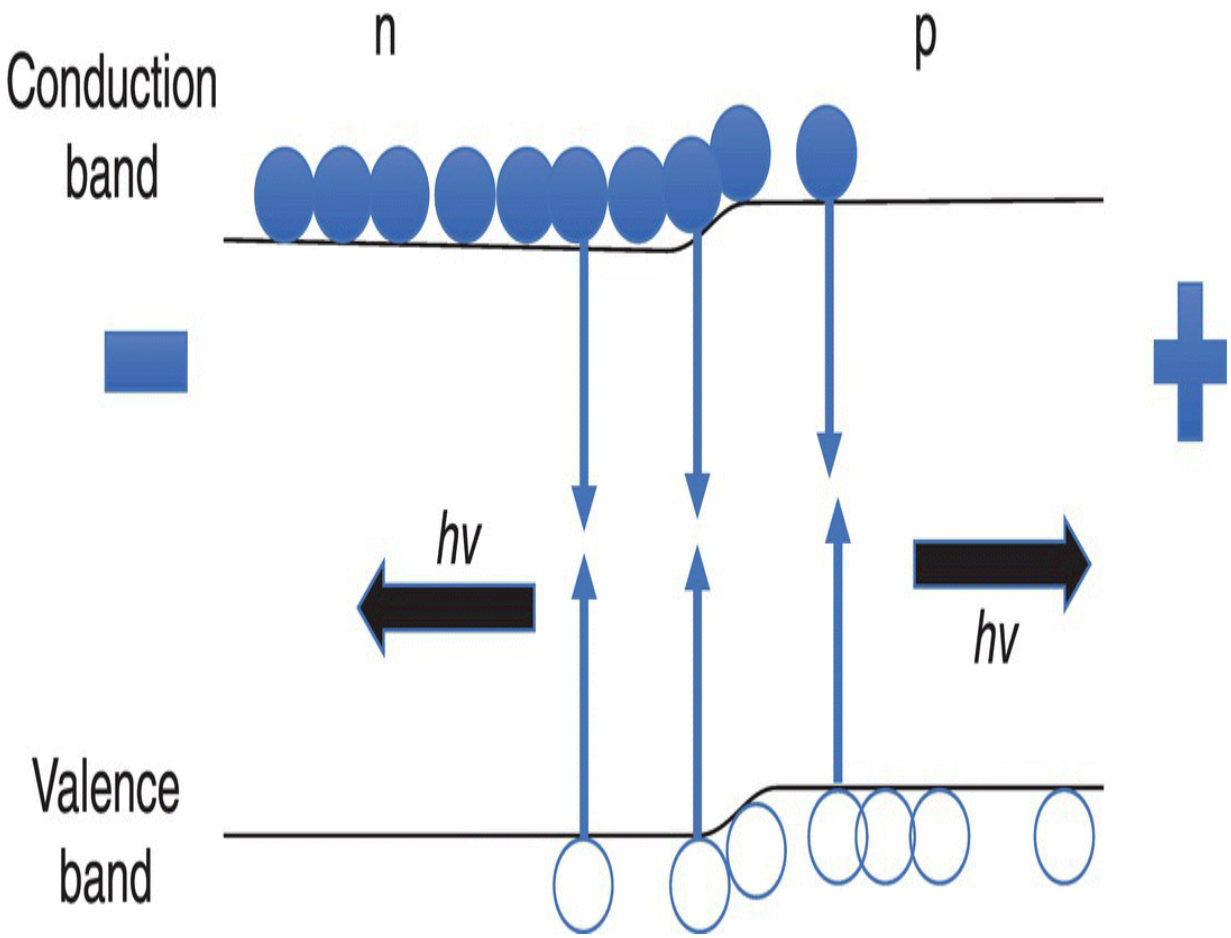


Figure 13.13 The spontaneous recombination of electrons and holes at the junction generates photons with a very specific frequency.

Figure 13.13 shows the already well known forward-biased diode. When we forward bias a pn-junction, as we have seen many times before, electrons from the n-type semiconductor move toward the p-type semiconductor and the holes move in the opposite direction. The electrons can now recombine with holes, as I show in Figure 13.13, and in the process generate light with an energy equal to *hv*. The frequency of the light depends on the energy gap of the semiconductor we use. The recombination occurs at the transition region or very close to it. We make the transition region of the p-type region very thin so that the photons generated are not reabsorbed into the bulk. We also make the n-type region very

heavily doped so the majority of the transition region falls on the p-side.

We have seen that radiation that comes from diodes has a fixed frequency. How do we get white light? This can be done in two ways. First, the diode can be made with several different semiconductor layers so the combination of two or three colors results in white light. The second method is to use a diode that emits ultraviolet radiation through a phosphor-coated window. The ultraviolet radiation strikes the phosphor coating and the phosphor re-emits the radiation in the visible light frequencies.

Table 13.1 list some of the semiconductors used to get the right color. The different ratios of arsenic (As) and phosphorous (P) result in different colors. To get white light we combine blue, green, and red. These LASERs and LEDs are devices that use compound semiconductors where Ga has a valence of 3 and both As and P have a valence of 5. This is why the proportion of As and P must equal 1 to create the perfect Zincblende structure (see Figure 3.4) needed for a good diode.

Table 13.1 LED semiconductor materials used to obtain different colors

Color/frequency (nm)	**Semiconductor used**
Blue/470	SIC
Green/550	GaP
Yellow/590	$GaAs_{0.15}P_{0.85}$
Red/650	$GaAs_{0.6}P_{0.4}$
Infrared/1000	GaAs

The advantages of LEDs are well known. They last as much as 100 000 hours (4000 days or 11 years) and are considerably more efficient than incandescent bulbs in which the majority of the energy is transformed into heat.

It is also interesting that there is a law, Haits law, which is similar to the Moore's law, that says that every decade LED performance (i.e. the flux) will increase by a factor of 20 and the cost will decrease by a factor of 10.

13.5 Summary and Conclusions

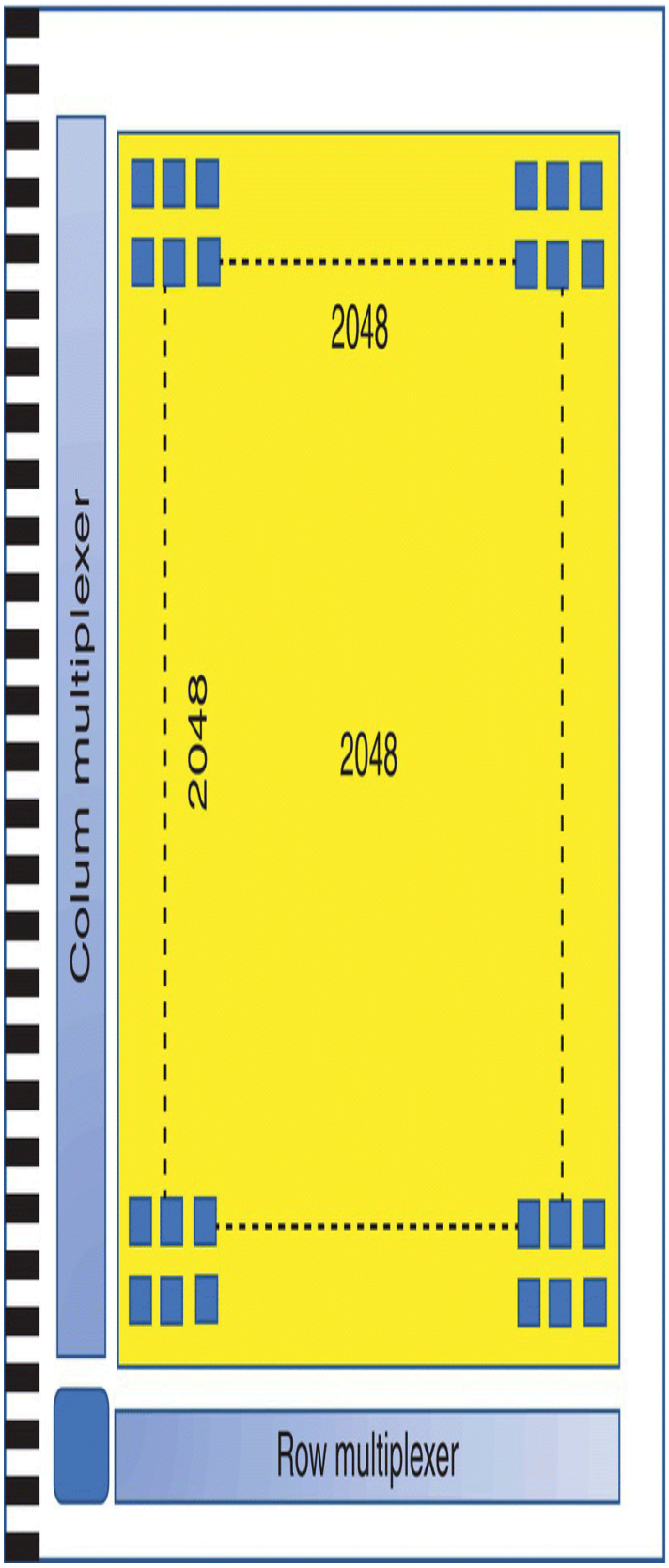
In this chapter we have covered several optoelectronic devices, that is, semiconductors that interact with photons of light. All depend on allowing photons of light to give enough energy to the semiconductor material to kick an electron from the valence to the conduction band. In this way we recognize that a photon has hit the semiconductor. In LASERs and LEDs, we use the opposite process, that is, electrons dropping from an excited energy level to the lower ground state give up the energy as light photons. In LASERs we need to have a population inversion, an exited state that has more electrons than it should, with cavities that bounce the light back and forth, stimulating more electron–hole pairs and thus obtaining a coherent light that grows in intensity. With semiconductor LASERs we have seen another use of what we call “degenerate” or highly doped semiconductors. In LEDs, although the physics is very similar to that of LASERs, we just let the photons shine. We use different semiconductor materials to obtain different light frequencies and therefore different colors. I again want to point out that some research that seemed esoteric and a waste of money at first resulted in a very useful device that has revolutionized a large number of technologies, from medicine to science to manufacturing to domestic uses.

Appendix 13.1 The Detector Readout

Toward the end of Section 4.5, after I explained how detectors work, I mentioned that I would explain the readouts, the circuit we use to readout the information detected by the detectors, after we have covered some of the components involved in the electronic design.

Now we can go back to this and understand the detector readout chip.

We left Chapter 4 with the detector array forming a matrix of 1024 × 1024 or 2048 × 2048 individual detectors with indium bumps ready to be connected to the readout. I show the detector array and the readout in Figure 4.14. We covered the indium bumps in Section 10.9. Each one of the millions of individual detectors needs its own reading input. We do not want to combine signals. That would not be a photo of what the infrared detector sees or what an object or the sky looks like. Figure 13.14 shows the structured of the readout array and the unit cell. On the left of Figure 13.14, I show the structure of a detector readout array. Each box in the larger area has the same size as the detector pixels. The detectors are read sequentially using two multiplexers (MUXs): one vertical that selects one row of electrons at the time and one horizontal that enables one of the columns. The MUXs drop the information for one cell at a time to an output amplifier, the lower left corner of the array, that sends the information to the data-processing computer. In many cases the contacts are all located on one or two sides of the array, as I show, so we can butt the arrays very close to each other to form a larger array.


Colum multiplexer
2048
2048
2048
Row multiplexer

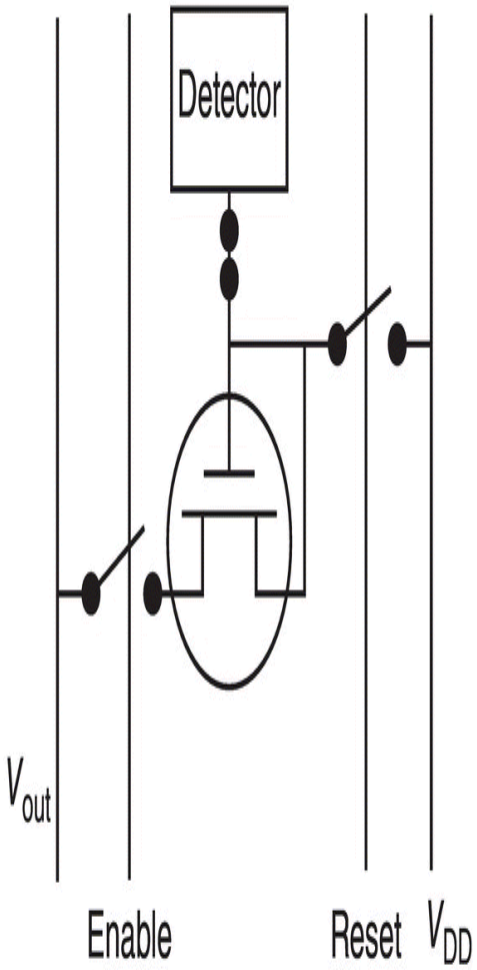

Detector
V_{out}
Enable
Reset
V_{DD}

Figure 13.14 A typical detector readout array with as many inputs as detector pixels, with MUXs to read out all the individual detectors one by one.

I show the unit cell of the readout array on the right of Figure 13.14. It is quite simple. It has one analog CMOS and two digital CMOSs, which I show as switches. The switches are normally open, isolating the CMOS from the bias voltage and the output line.

We start by turning ON the reset line and resetting all the CMOSs to the bias voltage, V_{DD}. We turn OFF the reset line, isolating the readout CMOS. As photons hit a given detector the gate of the analog CMOS keeps on charging for as long as we keep on looking (as long as 20 minutes for some observational applications). The CMOS is disconnected from any input until there is time to read it. The vertical MUX chooses a row and all the detector readouts on that row are connected to their own vertical (V_{out}) line. The horizontal MUX sequentially reads the voltage of each of the detectors in each row. The amplifier sends the information to the external electronics and computers to interpret and create the images. This process is done extremely fast compared to the observation time. After we have read the whole array, we reset the readout CMOS, turn OFF the reset switch, and start to integrate the incoming radiation again.

14
Microprocessors and Modern Electronics

OBJECTIVES OF THIS CHAPTER

In Chapters 11 and 12 we looked at a number of semiconductor circuits that accomplish a specific function. Now we are ready to put many of these components together and look at how computers work. As I have tried to do throughout this book I will, as much as possible, go inside the computer and look at the relation of its inner working with what we have learned about semiconductors and semiconductor devices.

I will also cover liquid crystal displays, which are found everywhere, and show that although the media is the liquid cells, how they operate is based also on semiconductor technology.

14.1 The Computer

14.1.1 Computer Architecture

In Figure 14.1 I show the basic components of a modern computer. You'll find that different books show different computer architectures. It depends very much on how many details one wants to include and how the different functions are combined. For example, the central processing unit (CPU) includes the control, the arithmetic unit, and the registers. Figure 14.1 is just one simple way of showing all the major components. It also shows the interconnections between these large boxes. Some of the interconnections are just data that we want to move from one place

to another and some are instructions (darker lines) that tell the CPU what to do.

In Section 12.5 we covered, in quite a bit of detail, memories. The memories box in Figure 14.1 includes the long-term memory as well as the different cache memories that interact quickly with the arithmetic unit. We also covered many of the functions of the arithmetic unit in Chapters 11 and 12. Let's talk first about the three boxes outside the CPU and how they work in a computer.

The input box accepts the data from a keyboard, mouse, disk, finger on a screen or applications and data from the internet. The input unit has its own electronics that translates all of the different inputs into…” ‟ and the input box has its own dedicated microprocessor to perform this analog to digital conversion. I explain microprocessors in Section 14.2.

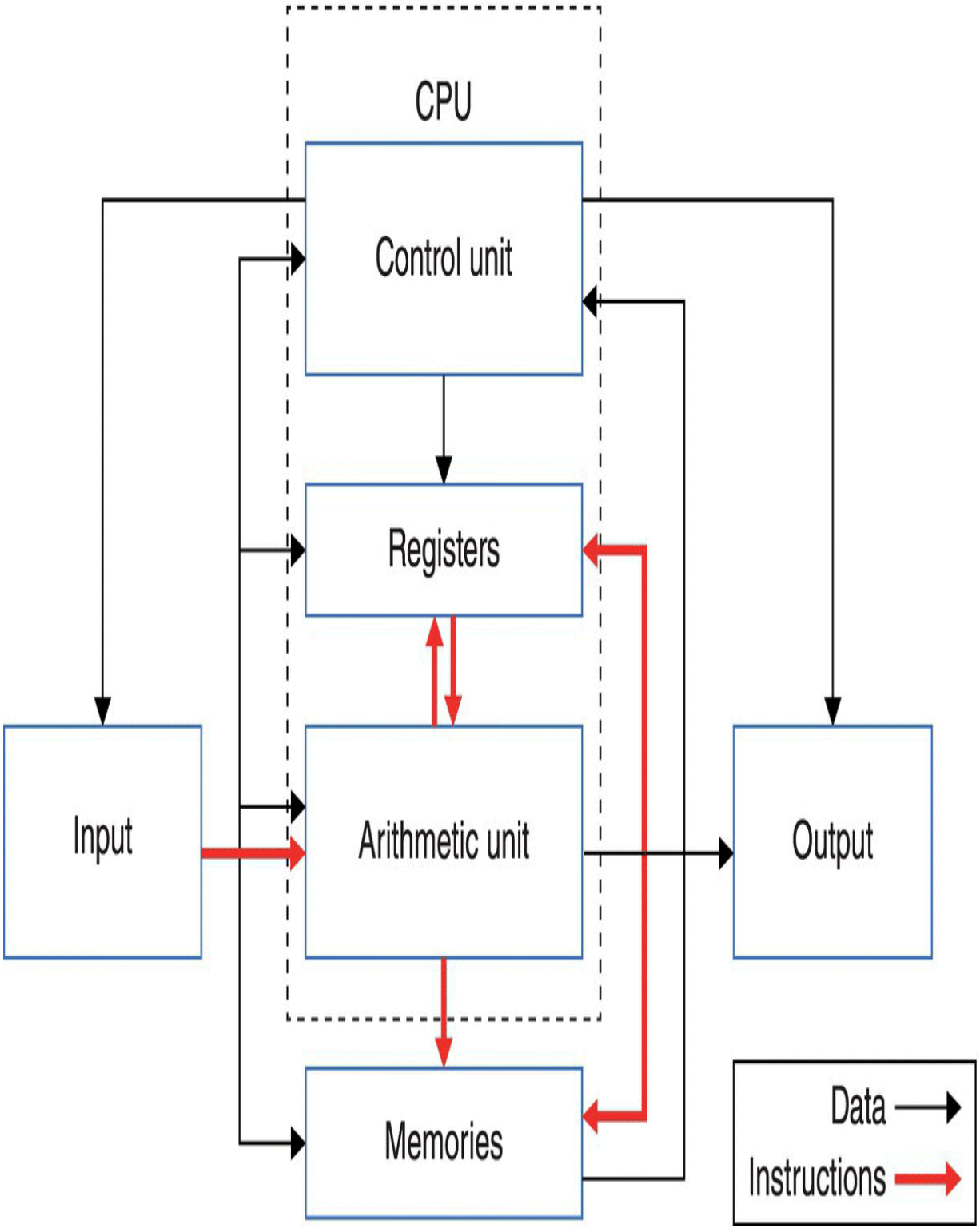


Figure 14.1 Basic components and interconnections of a modern computer.

The output box does the opposite. It takes the 1s and 0s that the arithmetic unit has calculated and stored in the memory and the output box translates the digital numbers into something we can understand and send to a printer, a screen or maybe just an external memory or external drive.

Other important components are all the arrows going from one place to another. These are called busses. Typically, we distinguish between the data bus and the address or instruction bus. The address bus tells the CPU where to find the information the arithmetic unit needs, and the data bus just takes whatever information is stored in a location and sends it to the appropriate register. These buses, like the motherboard where all the components are located and interconnected, are one of the more important concerns of a computer designer. One of these buses connects the computer to other external devices and it is probably the one you are familiar with, the USB cable (USB stands for universal serial bus). All devices, TVs, computers, cell phones, disk players, etc. have to follow the same specifications so they can talk to each other. These cables cannot transmit all the data at the same time. USB cables only have four wires: one is for power (5 V), one is the ground (0 V), and there are two cables for the data, one positive and the other negative. The two data cables are twisted to minimize or cancel any noise pick-up, therefore at both ends of the cable we need to have a multiplexer and demultiplexer, which I explained in Sections 12.1 and 12.2, and they determine in which order the data is read or written.

I should also mention that there is a timer that synchronizes all the operations. I discussed the timer, the timing, and the different wave forms in Section 12.4.

14.1.2 Memories

In Section 12.5 I discussed memories as independent components, the different types of memories, how they are physically implemented. and how they work. Here I talk more about how they

interact with the computer. The memory contains not only the information, the data, we have stored but also its addresses. This is quite important since the computer needs to find where the information it needs is stored. The control unit keep track of all these addresses.

The address space in the memory has to hold as many bits as we need to uniquely identify the location of the data or the information we need. [A bit is a "binary digit," usually written as small b, that has a single value of 1 or 0. Eight (8) bits is a *byte*, usually written with a capital B. The choice of 8 bits to a byte, not 7 or 9, is because 8 bits are sufficient to identify all the letters, numbers, and symbols of the standard keyboard (see Appendix 14.1). Thus, we use 1 byte every time we press a key. A "word" is 4 bytes or 32 bits.] Since each bit can have two numbers, 1 or 0, the number of addresses is equal to 2 raised to the *N*th power, 2^N. If I assign 16 bits for the address, I have a memory with 65 536 memory addresses and if I double that to 32, I get 4 294 967 296 or 4.3 GB. And every time we add another bit, I double the number of addresses. This also means that the address bus must contain 16 or 32 bits. This is known as the *width* of the bus. The number of bits I need for the address is usually larger than the number of bits that contain the information stored in that location.

The CPU is always much faster than the memories. We change the size and the speed of the memories as they get closer to the CPU. Consider the memory stack in your laptop (Figure 14.2). As we go away from the CPU, we have many (32 or 64) registers, three cache memories, levels 1, 2, and 3, the main memory, and finally the storage memory. As we go further away from the arithmetic unit, the memory size increases and the speed for accessing the data decreases. Outside the memories, not shown in Figure 14.2, are the tapes, CDs, magnetic discs, external drives, and memory sticks all located outside the computer itself (CD readers are outside the "computer" even though they may be located inside the box we call a computer). There are two main reasons for dividing the memory into so many levels. One is cost. Faster memories are much more

expensive per bit than slower memories. The second reason is power. The faster the memory, the more power you need to switch them on and off (see Section 11.14).

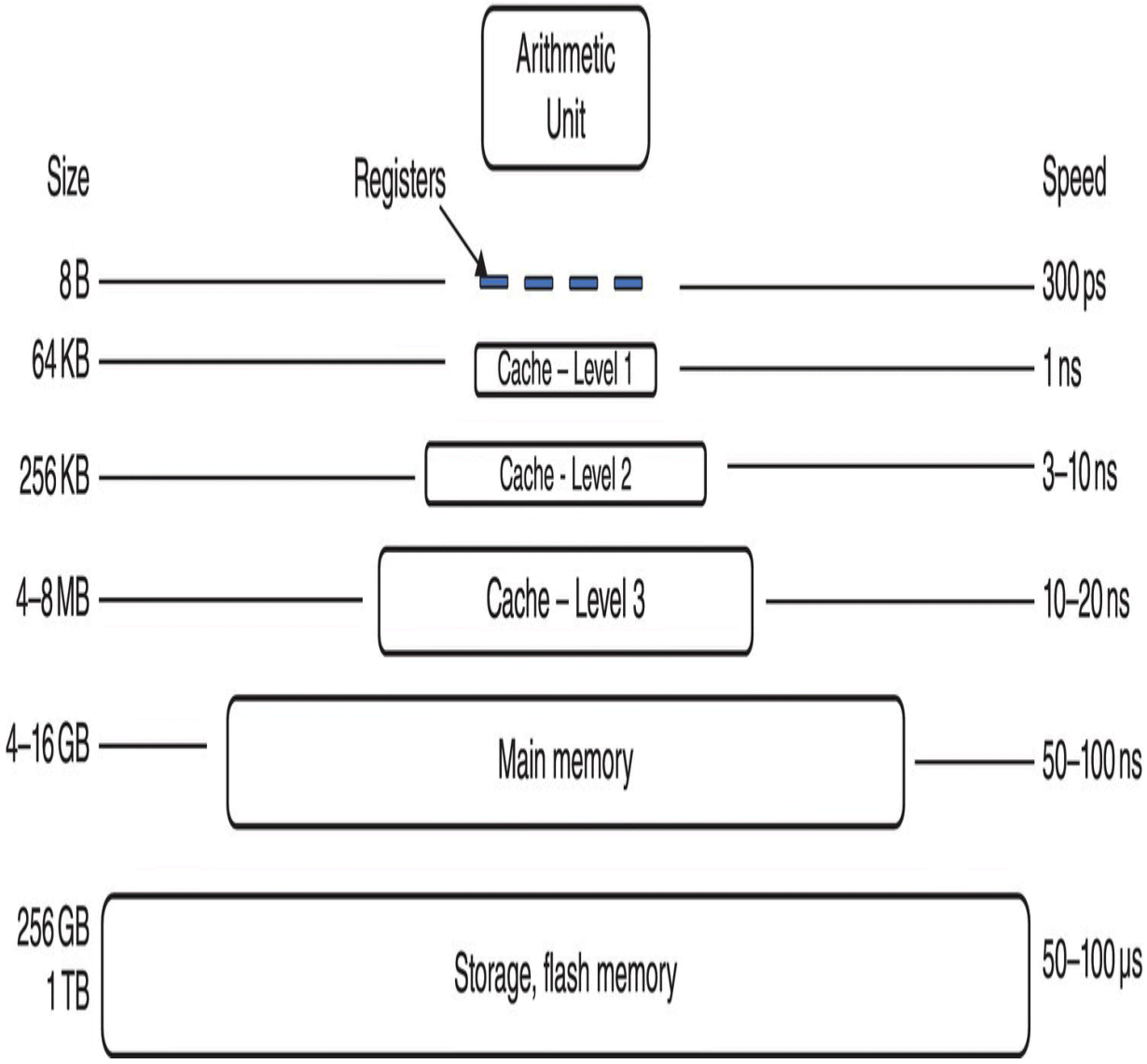


Figure 14.2 Memories in a typical laptop. The closer the memories are to the arithmetic unit, the faster their speed but the smaller the size.

An integral part of the CPU is the registers, which in a sense are memories that contain the information that the arithmetic unit needs right at that moment to perform a desired operation. If you want to add 2 + 3, the 2 will be in one register, the 3 in another, the

instructions to add them in a third one, and we need a fourth one to enter the result. I discuss registers in Section 12.3. A good computer has at least 32 registers. There is nothing that slows down the arithmetic logic unit (ALU) more than running out of registers. The register uses custom CMOS devices and has multiple inputs and outputs to perform very fast operations which can typically be addressed in 0.1–0.3 ns.

The next level of memory is cache memories. Their size goes from 64 kB to 8 MB and the access speed decreases accordingly from 1 to 20 ns. These memories use SRAMs (which I discussed in Section 12.5.1) because they are faster.

Large memories use DRAMs (Section 12.5.2). DRAMs can be constructed with just one MOSFET so the packing of units, bits, is considerable tighter and we can have more bits in a chip. The DRAM is a slower memory for many reasons. We need to select the row and the column that stores the information we want. A 64 GB memory has over 25 000 rows and columns. They have to be selected by multiplexers (Section 12.1), which takes time. Furthermore, the information on the DRAM has to be refreshed or written back as soon as the data is read. The access time is about 100 ns.

Finally, the largest memory, the one that holds all my 50 000+ photos, the 60 or more applications, all the documents, letters, and reports I have written and all those I have received in the past 30+ years, is the disk storage, which can take up to milliseconds to retrieve the required information. Flash memory is actually an EEPROM, which I mentioned in Section 12.5.4.

14.1.3 Input and Output Units

The input and output units connect the computer to the outside world. Most requests from the control unit to the memory and ALU are internal operations that move data from one memory to another or from a memory to a register or from one register to another. At some point, we need to extract information from the computer and

read it on a screen or print it using a printer. The input/output (I/O) unit also includes a graphic card, a serial/firewire to interphase with the disk, a keyboard, a mouse, modems, other computers, your digital camera, etc.

Microprocessors work with parallel data, so there is an I/O controller that changes the serial inputs (remember the USB cables have only two data lines) into a parallel data stream to interphase the I/O to the microprocessor.

I should also remind you that, as far as the computer is concerned, there is no difference between "data" and "instructions." For the computer it is all 1s and 0s. and all of them share the same memories and registers. The address is what tells the computer where the desired content of the memory is located.

Buffer memories are temporary memories used to store data or information on the way from one location to another. Sometimes the information comes in faster than it goes out. The buffers hold data when it flows in and out at different rates.

14.1.4 The Central Processing Unit

As I show in Figure 14.1, the CPU consists mainly of three components: the control unit, the ALU, and the registers. The CPU tells every other part of the computer what to do and when to do it. The instructions for the control unit are also stored in some portion of the memory but the control unit tells the memory what instructions to look at. If you want to add two numbers, for example, the control unit goes to the right memory location and requests that the two numbers go to the desired register(s) in the ALU. It tells the arithmetic unit to pick the numbers from such and such a location in the register, add them, and send the result to a location in another register. At some point it directs that information to the output unit and displays the result in any one of the output devices. These are the four elements of an instruction:

DO THIS – operational code, also known as opcode

TO THIS – operant in such and such a location

PUT IT THERE – location address of the result

GO TO NEXT – when you finish.

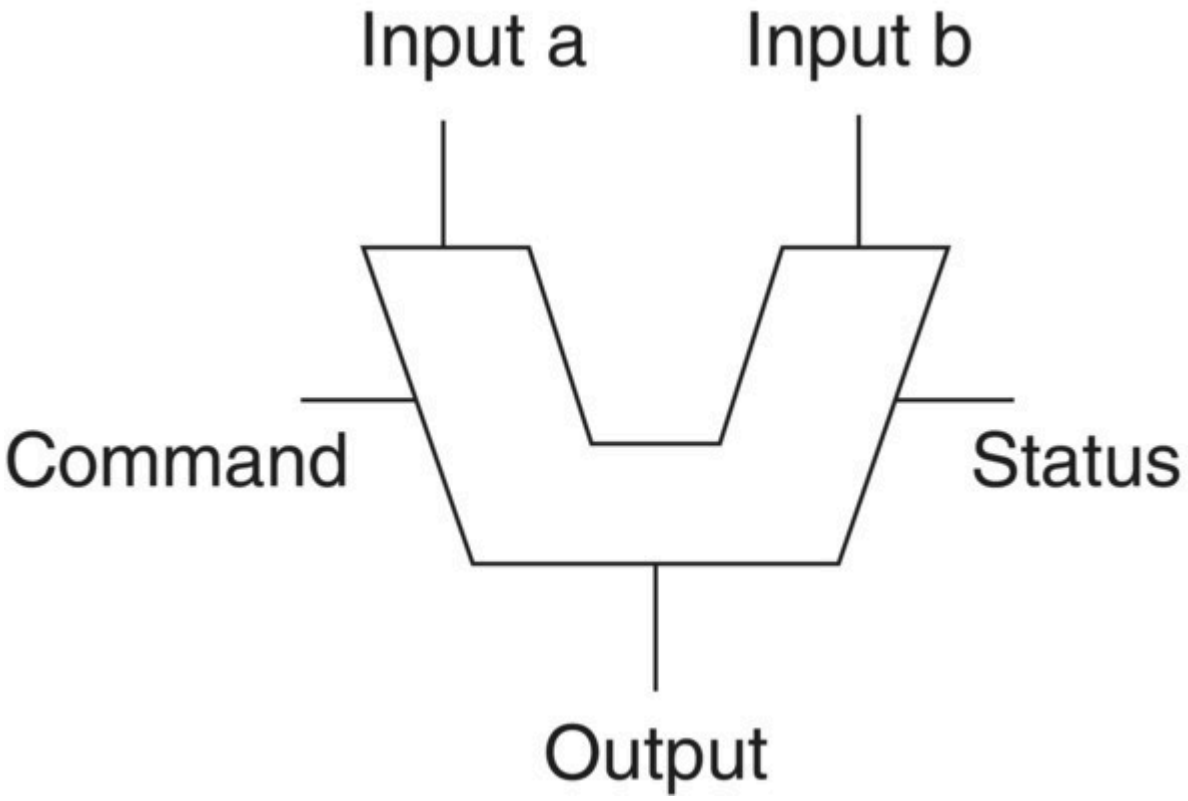


Figure 14.3 Symbol for the ALU.

Finally, the most obvious component of the computer is the ALU. I show the symbol for the ALU in Figure 14.3.

The ALU has two (or more) inputs. Let's say, for example, that input a is a 2 and input b is a 3. The command input maybe the "add" instruction, and the output is the result, the number 5. All of these, obviously, are in digital notation. The ALU has added the two numbers. The status tells the ALU of any conditions that may influence the result, such as if the result is positive or negative, or if I want only the absolute value or how many numbers after the decimal point I want. The ALU has many more than just two inputs and one output. Generally, the ALU will have 64 inputs and 64 outputs that means also that data busses will be bundles of at least

64 wires to send the data in parallel, much faster than if each bit had to be received or sent one at a time.

The CPU needs instructions to operate, and these instructions must be in machine language, that is, binary code. The instruction set consists of arithmetic or operational code, opcode (add, subtract, load, store, etc.), logic (AND, OR, etc.), and others such as shifting or rotate instructions. It also needs to know the addresses where all this information is stored, the source operand code, and where to place the result, the result operand code.

I show in Figure 14.4 in a little more detail on how the CPU performs an operation. The CPU starts the process by first looking to see where the address of the instruction it has to perform is located. Once it knows where it is, it gets the instructions and decodes them so the ALU can understand them. Next it looks for the address of the operands and gets it. There is always more than one operant, so the CPU has to do this operation a few times. Once it has the data and the instruction to do the desired operation, the ALU goes ahead and does whatever it has been asked to do. The last thing the CPU does is find the address where the result needs to be stored and send the result there. The operation is completed so now the control unit tells the CPU what the next instruction is, and the process starts over again.

The CPU does all the steps I show in Figure 14.4 in parallel. After it reads the first instruction (second box) at time t_1, it sends it to the decoding box (third box). At the next time period, t_2, the CPU not only decodes the information and sends it to the fourth box, but it also searches for the next instruction while decoding the first one. So, two boxes are now running simultaneously. At next time, t_3, a third instruction comes up, the second instruction is being decoded, and the first instruction is now finding the operands. All the boxes are operating simultaneously with successive instructions and operations. We call this *pipelining*. It is very similar to Ford's assembly line: we do not want any station to be idle.

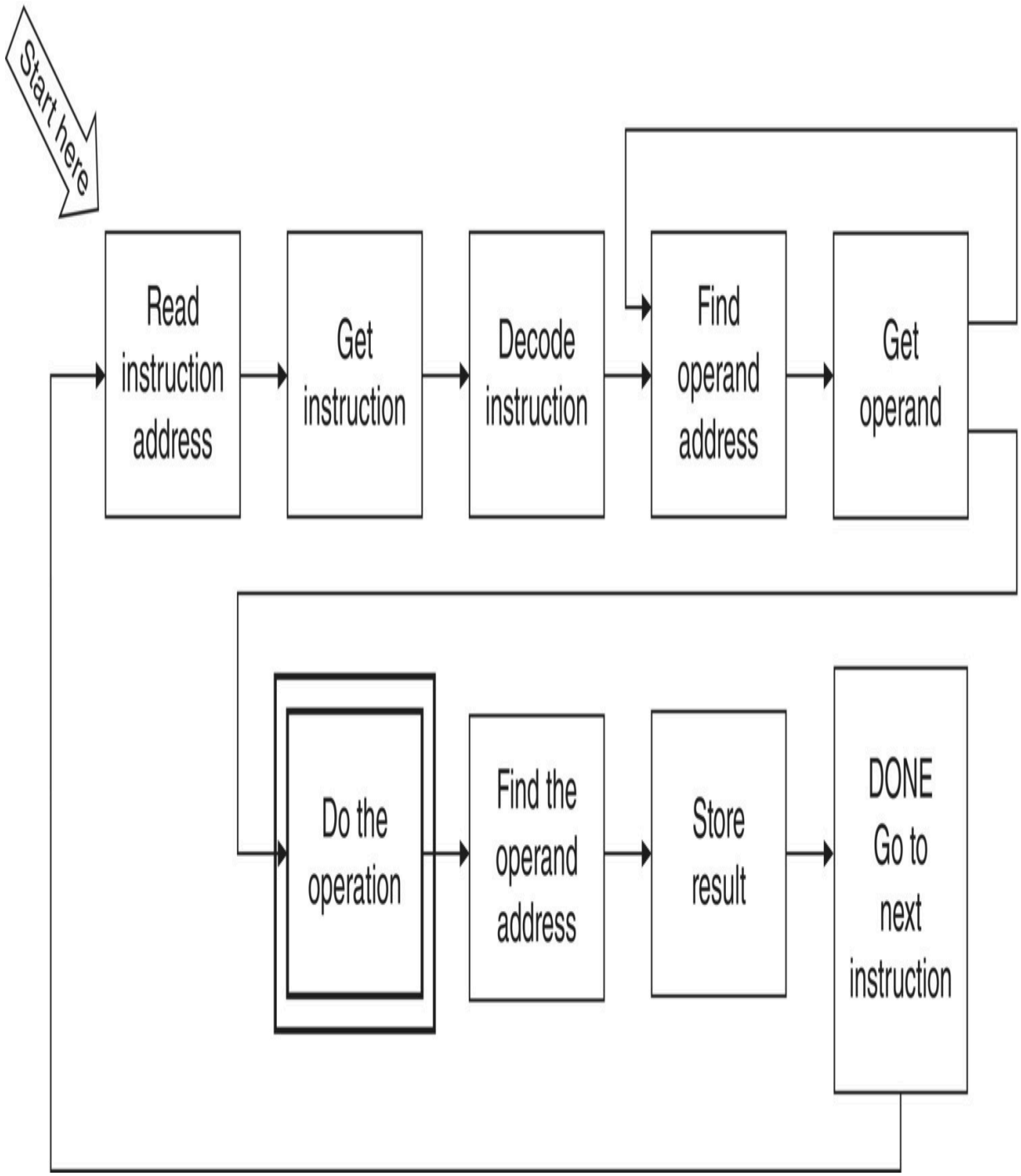


Figure 14.4 The CPU processes an operation sequentially and when it finishes it looks for the next thing to do.

Speed is very important. One of the ways to determine the speed of a computer is the MIPS value. MIPS stands for millions of

instructions per second. The intel core is up to 250 000 MIPS. You can figure out how fast this is.

14.2 Microcontrollers

Microcontrollers are basically less powerful, smaller computers dedicated to a specific task. Microcontrollers are used in cars, TV sets, washing machines, card readers, telephones, and myriad consumer products, almost everything that has some sort of automatic operation, except computers themselves, although inside a computer there are many microcontrollers that do very specific tasks like interpreting the information coming from the keyboard and translating the information into machine language. These are called *embedded* systems since they are an integral part of the product you buy. Many products, like cars, for example, contain tens of microcontrollers dedicated to a particular function. A microcontroller in a smart thermostat is able to store the present date and time, the temperature you want, the time when the temperature has to change, order the heater or the air-conditioning unit to turn ON or OFF, when to turn the fan ON, and perform many other actions, but it cannot add 1 + 1.

The microcontroller contains in a single chip the CPU, the memory, both ROM and RAM, and the I/O control functions. It may contain other functions such as analog to digital converters. Many microprocessors are programable so different companies can buy the same processor and force it, by programing, to do a specific task(s). In fact, microprocessors are complete computers but with limited capabilities. They are also quite cheap, if bought in large quantities.

14.3 Liquid Crystal Displays

Liquid crystal display (LCDs) are everywhere. The TV, of course, but also your phone, your computer screen, video games, watches, and all types of instruments. The advantages of LCDs are many: image sharpness, excellent brightness, very little optical distortion, low

power, and low weight. But why do I discuss them here at all? They are not semiconductor devices.

Well, let's see. I show the elements of an LCD in Figure 14.5. The reason I cover LCDs is because one of their main and necessary components is thin film transistors (TFTs) (in violet). I will come back to this later but first I want to explain all the components of an LCD screen.

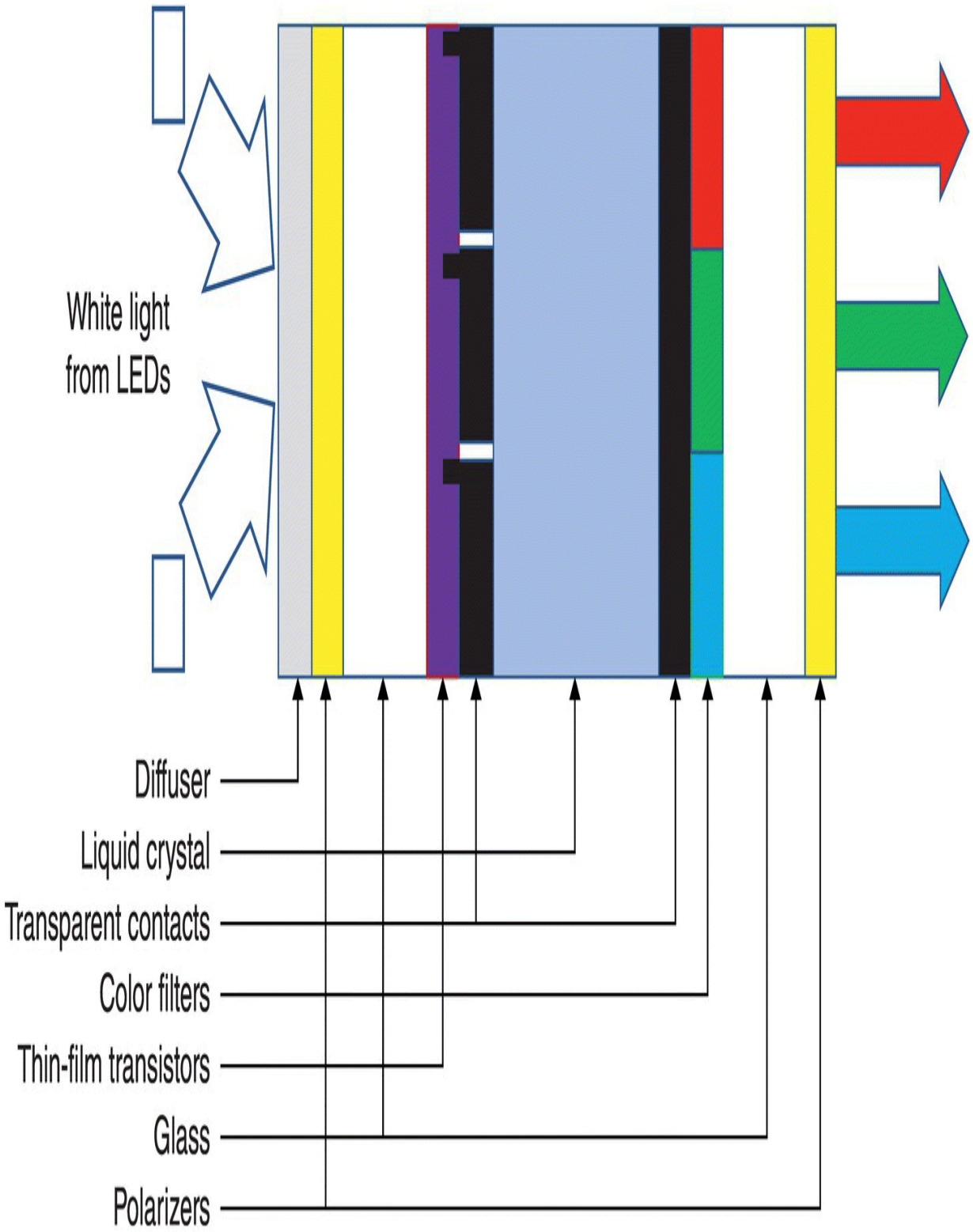

White light
from LEDs
Diffuser
Liquid crystal
Transparent contacts
Color filters
Thin-film transistors
Glass
Polarizers

Figure 14.5 The main components of an LCD. The liquid crystal is in the middle (light gray), sandwiched between two transparent contacts (black), the first one pixelated, a color filter (multicolor), a TFT matrix (violet), two glasses (white), and two polarizers (yellow).

14.3.1 Liquid Crystal Materials

Liquid crystal material is an organic compound composed of very elongated molecules. The ratio of the length to the width of one of these molecules can be as large as 7 to 1. Figure 14.6 shows one such molecule. Depending on the temperature, this chemical material has three phases (Figure 14.7). At very cold temperatures, 0 °C in Figure 14.7, the liquid crystal is solid (like ice). All the molecules fall in line. At high temperatures, 50 °C or higher in Figure 14.7, the molecules are not oriented, they are arranged randomly any which way as they would in any other liquid. At in-between temperatures, the molecules start separating from each other like in a liquid, but still conserve their original orientation. This is the liquid crystal phase (I wonder why we do not call it semisolid or semiliquid as we do with semiconductors!). If we coat the plates that encapsulate the LED molecules with a thin alignment layer, the molecules orient themselves with the layer orientation (Figure 14.8).

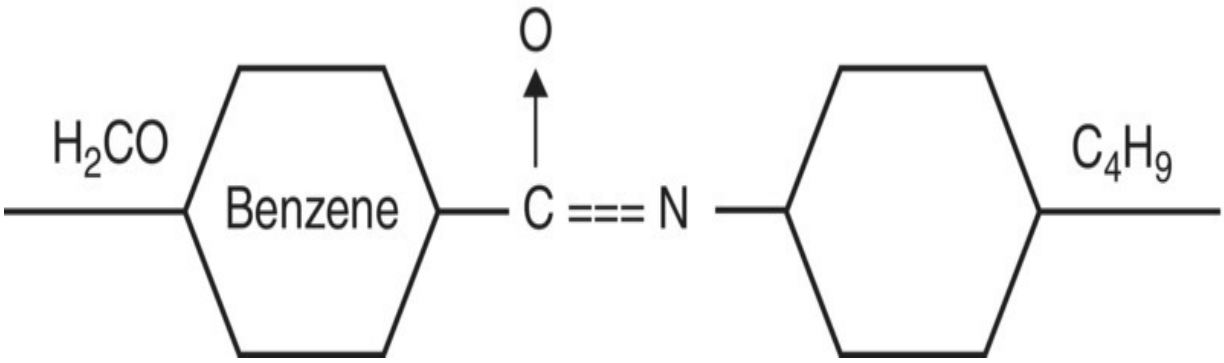


Figure 14.6 Molecule of a liquid crystal consisting of two hexagonal benzene molecules holding hands with a carbon and a nitrogen atom. Its chemical name is a mouthful, methoxybenzylidene.

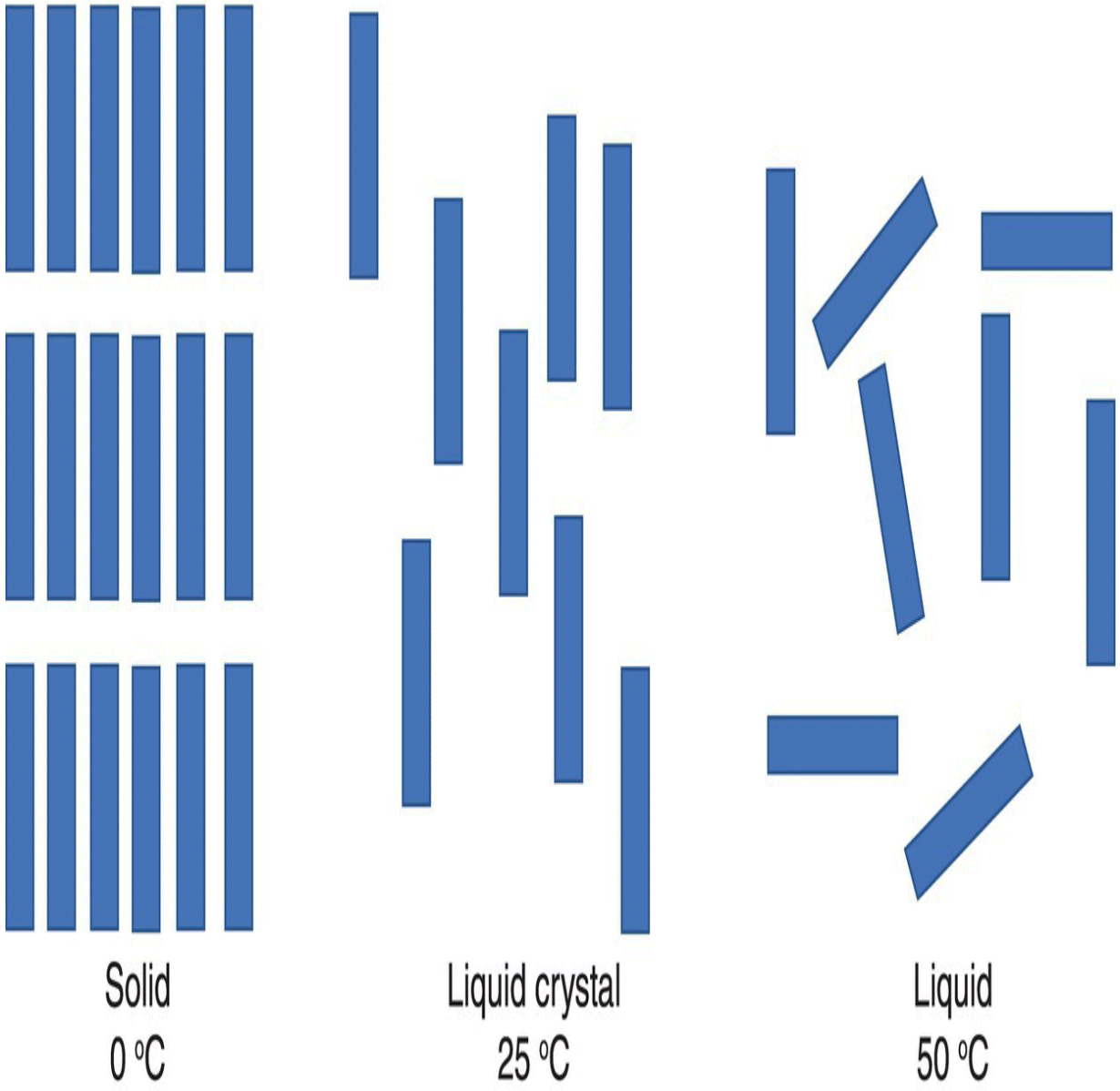


Figure 14.7 The three phases of a liquid crystal: solid at 0 °C, liquid crystal at intermediate temperature, and liquid at 50 °C.

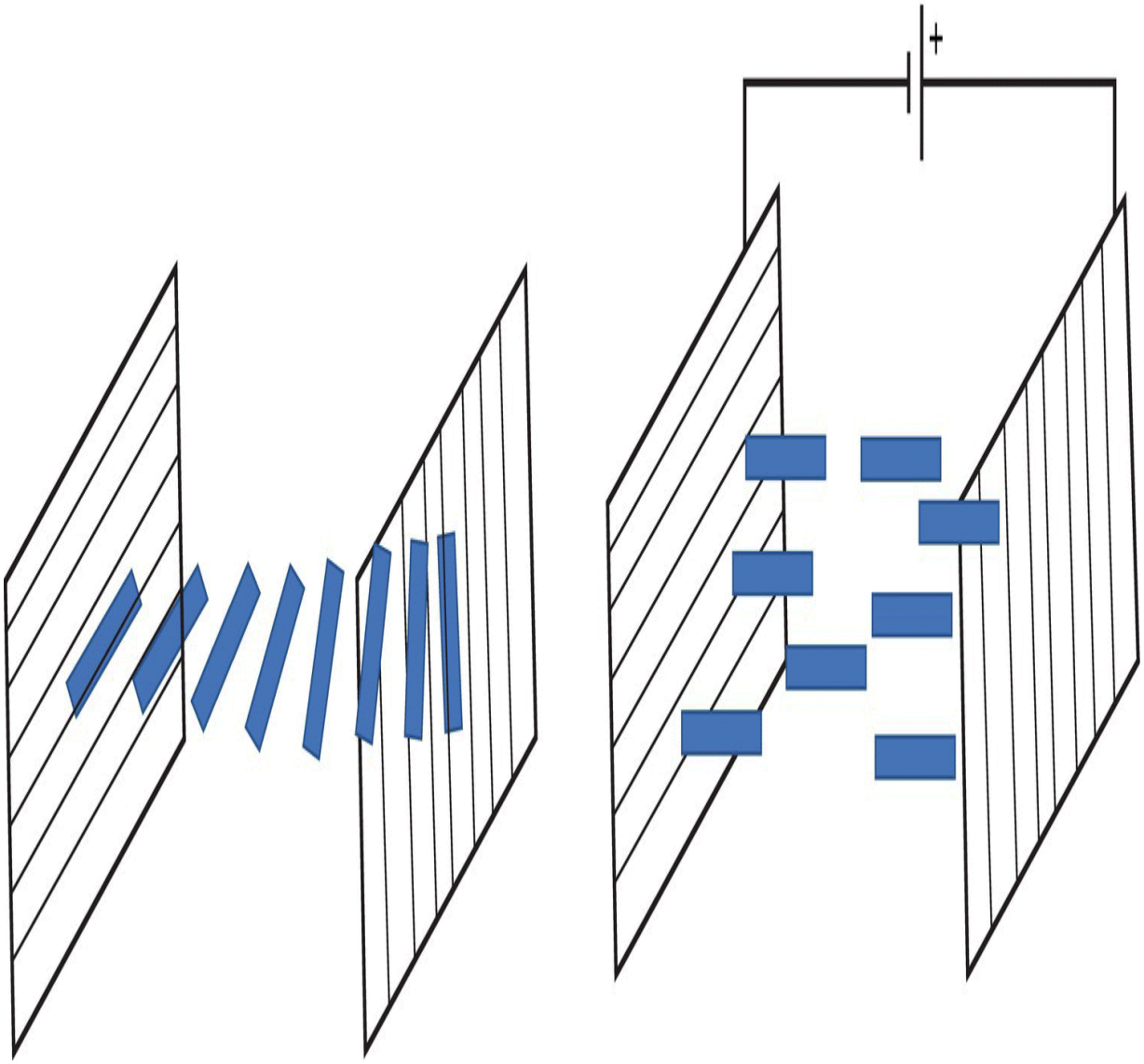


Figure 14.8 The liquid crystal molecules align themselves with the two contact layers that align at 90° to each other (on the left). If we apply an electric field (on the right), we force them to alight themselves in one direction.

The molecules touching the contacts must have the same orientation as the alignment layers in the contacts. If the alignment in one contact is rotated 90° with respect to the other, as I show on the left of Figure 14.8, the molecules are forced to align horizontally at the left surface and vertically at the right. In between, their orientation changes smoothly from one orientation to the other, from horizontal to vertical, so they can satisfy both boundary conditions. When we

apply a voltage, as I show on the right of Figure 14.8, we generate an electric field between the two contacts that forces the molecules to orient themselves in just one direction. When the light shines on the left of Figure 14.8 photons traveling through the liquid crystal are scattered by the molecules oriented in all directions. On the right, the molecules are all aligned and the photons are not scattered and go straight through the liquid media unchanged.

14.3.2 Contacts

Now that we know about the behavior of liquid crystal molecules let's examine the contacts. The most important properties of the contacts are, first, they have to be transparent so as much of the visible light as possible can go through the media. One of the semiconductors used as a contact is indium-tin-oxide, ITO. The proper combination of these elements results in a semiconductor with an energy gap of around 3.6 eV. The highest energy of visible light is 3.3 eV, which means that this semiconductor compound does not absorb any of the photons and is transparent to all visible radiation. (You can see the many uses we have of semiconductors and how the energy bands explain many of its properties. This is one reason why LCDs have a place in a book about semiconductors). Additionally, we make the contact very thin, which is the second typical design consideration. The contact has to be conductive, therefore we like to have high doping, but the higher the doping the less transparent it becomes. Similarly, the conductivity of the contact improves with thickness, but the thicker we make it, the lower the light transmission.

On the side of the optical filter the contact is a continuous sheet, but on the side of the thin semiconductor the contacts are small squares that define the dimensions of each pixel (see Figure 14.10).

14.3.3 Color Filters

The next component is color filters (see Figure 14.5). There is not much to say except that each filter absorbs the colors we do not

want. After we deposit a dyed gelatin at the other side of the contact, using photolithographic techniques similar to those I explained in Section 10.5, we open one set of holes on the photoresist and deposit one of the colored dyes. We repeat this process three times to deposit filters that allow red, green, and blue light to pass through.

14.3.4 Thin-film Transistors

Now comes the relevant part, the semiconductor. First a couple of words about thin-film transistor (TFT) technology.

I have mentioned several times before the need for single crystal silicon with no imperfections or impurities to fabricate the best transistors. I have also mentioned that we are able to fabricate 300 mm diameter wafers. We are not yet able to fabricate 5 ft diameter boules. If we needed them, they would be fabulously expensive. This is where TFT technology helps.

The TFT is basically a thin layer of polysilicon material, polycrystalline silicon, deposited on top of glass. This polysilicon is pure enough to fabricate CMOS transistors that are used only as switches. These polysilicon layers can be doped so we can create n- or p-type material and degenerate regions for the contacts and the metal lines that connect each MOSFET to the electronics. We can grow oxides and metal layers on top of them.

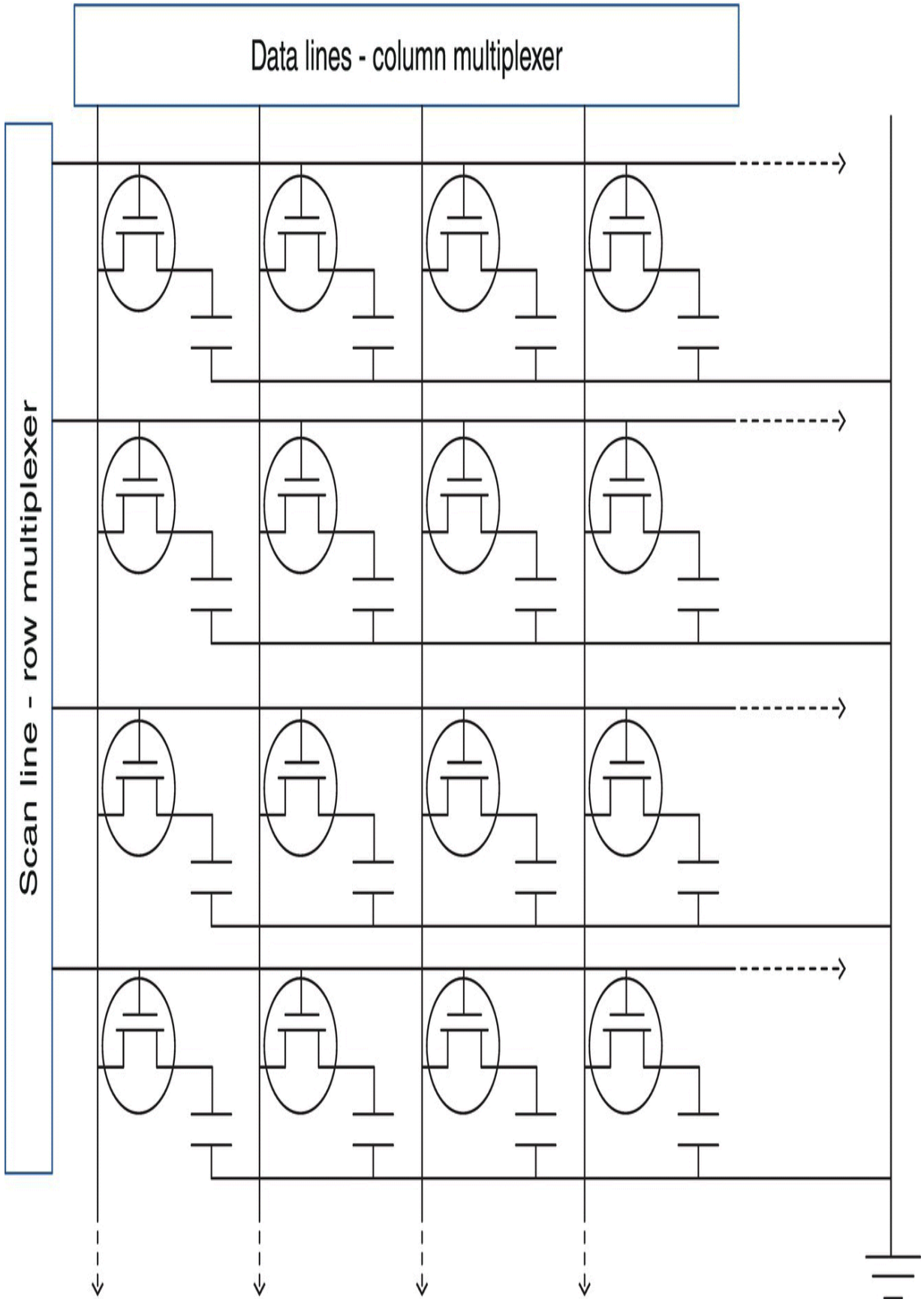

Data lines - column multiplexer
Scan line - row multiplexer

Figure 14.9 A partial matrix of CMOS switches that turn ON and OFF each of the liquid crystal pixels.

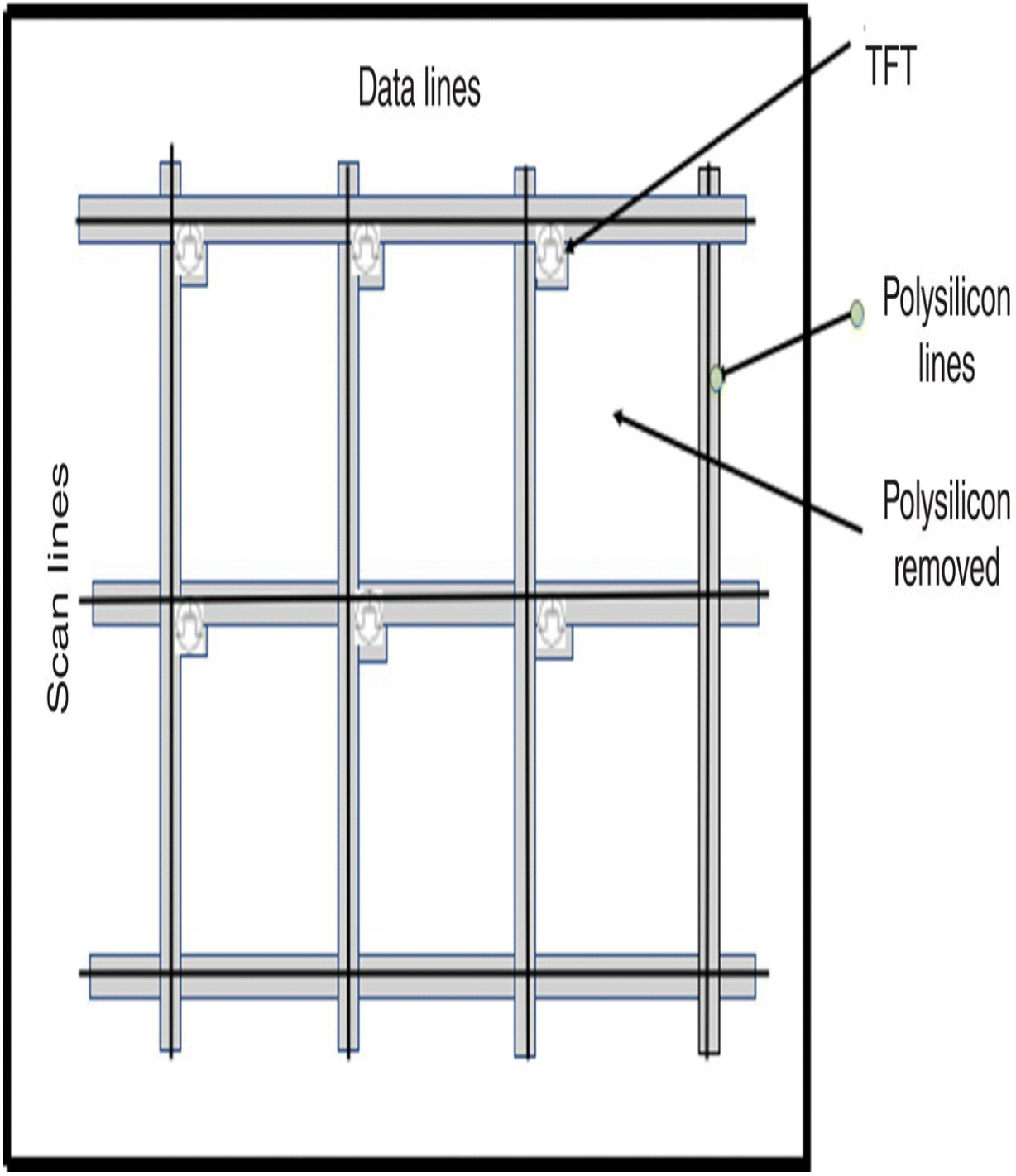


Figure 14.10 A partial array showing, not to scale, the portions of the polysilicon layer that are etched away to make the contacts more transparent. The contacts themselves are the capacitances I show in Figure 14.9 and the isolated contacts define the size of the pixel.

The unit cell is a very simple switch with a capacitor connected to the ground (Figure 14.9). These unit cells are arranged as a matrix with two multiplexers, one vertical for the data lines and another horizontal for the control or scan lines, just as we saw for detector readouts in Appendix 13.1. All the gates of the CMOS are connected to the row multiplexer. All the sources are connected to the vertical lines, the data lines. All the drains are connected through a capacitor to ground. We saw something similar when we talked about memory arrays. When a CMOS switch is ON, the capacitor charges to the voltage of its data line. When we select one scan line, then all the CMOS switches in this row are ON, and the capacitors in the row are charged to whatever value is coming from each of the data lines.

The LCD pixels are huge compared to the size of the transistors. The transistors and the polysilicon and metallic connecting lines occupy maybe 1 or 2% of the entire pixel area, therefore much of the unused polysilicon is etched away and removed so that there is better light transmission (Figure 14.10). The ground is connected to the other side of the capacitor.

A couple of numbers. A home LCD TV has an area of between 6 and 20 square feet, or between 500 000 and 2 million square millimeters. It also has 1920 vertical lines and 1080 horizontal lines for a total of 2 million pixels. Therefore, the size of each pixel is between a quarter and 1 mm^2. Transistors are about 10 000 times smaller, so you can see how much empty space there is. (if you look closely at a large TV screen without even using a magnifying glass, you will be able to distinguish the individual pixels).

14.3.5 The Glass

There is little to say about the glass (See Figure 14.5). The purpose of the two glass panels is to support, hold together in place, all the layers of coatings, contacts, and filters and the thin film material between them. Let me use this subsection to talk about dimensions.

The liquid crystal is typically 5 μm thin (a hair is typically 50 μm, about 10 times thicker). This very thin separation between two large glasses is accomplished using spacers, either just small transparent 5-μm balls or fibers of the same size. The glass itself may be 1–5 mm thick if necessary to support the entire device. The thin polysilicon is about 1 μm thick.

14.3.6 Polarizers

Light is an electromagnetic radiation. It has an electric field in all directions, as I show at the bottom of Figure 14.11. The radiation then has two components, x and y. The polarizer is a transparent plate that that lets only one of the light components, either x or y go through. You probably have a pair of polarized sunglasses. They are polarized vertically because most of the illumination we receive, from the sky and reflections from surfaces, are horizontal. Figure 14.11 shows two polarizers, A and B. As light passes through the first polarizer only the electromagnetic components parallel to the direction of the polarizer go through. Since electromagnetic radiation has two components, vertical and horizontal, when light passes through a polarizer it loses one of these components, decreasing the amplitude by a factor of two so only 50% of the light goes through the first polarizer. If we put both polarizers in the same direction, as I show at the Figure 14.11b, the polarized light from the first polarizer goes through the second polarizer without a problem. When we rotate the second polarizer (Figure 14.11c), the magnitude of the light decreases and if the second polarizer is rotated 90° (Figure 14.11d), then the light intensity goes for all practical purposes to zero. In 1820 Etienne-Louis Malus (1775–1812) discovered this phenomenon by looking at the reflection of sunlight

as it passed through the calcite windows of the Luxemburg palace in Paris. (Those were the good old days when scientists were treated like royalty.)

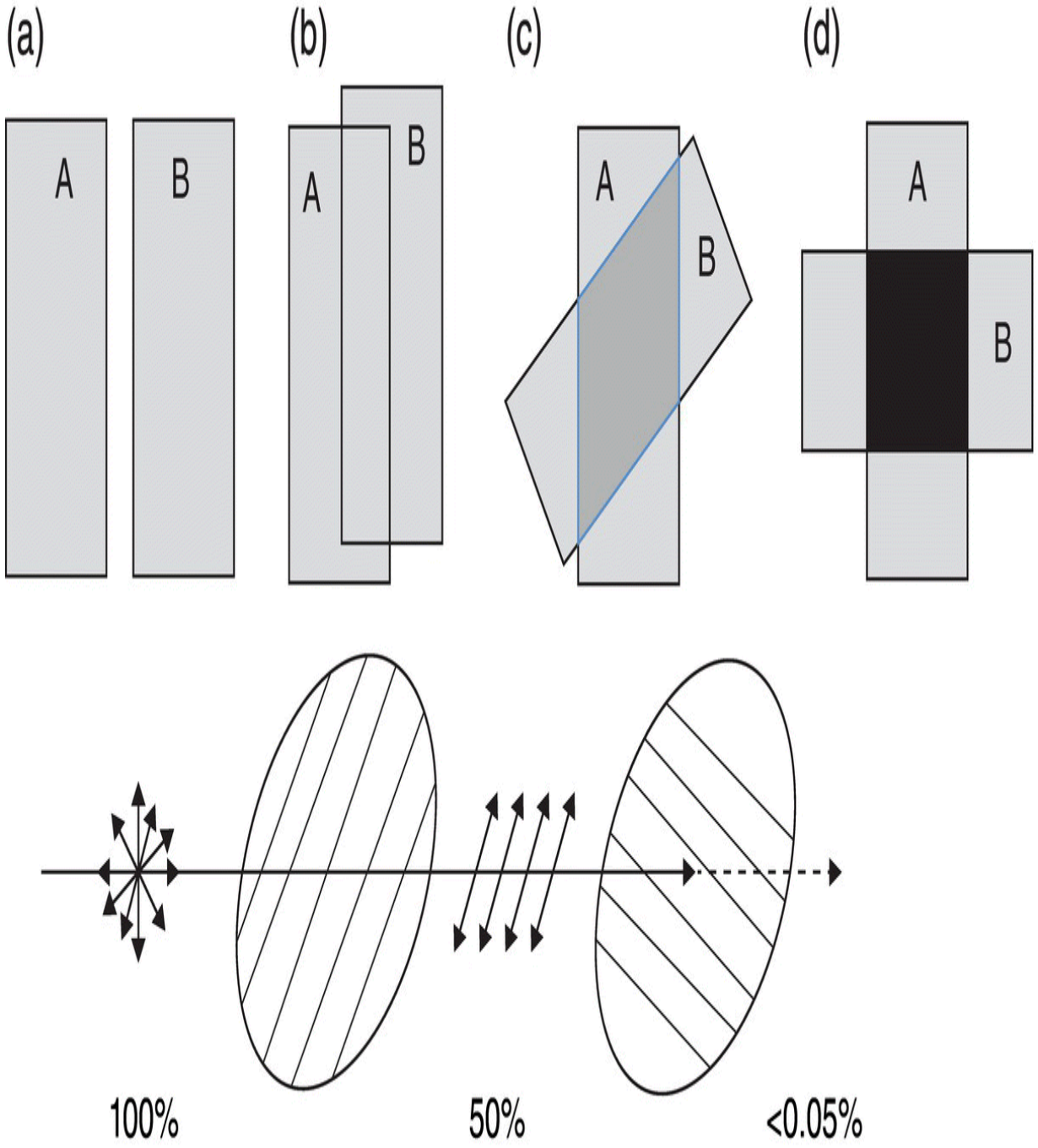


Figure 14.11 Top: a pair of polarizers, A and B. Both are polarized in the same direction (a and b). When rotated 90° (d) the light does not transmit through the second polarizer. At 45° the amplitude of the light decreases (c). Bottom: the electrical vibrations of the light as it goes through two polarizers rotated 90°. In between the light vibrates in only one direction.

14.3.7 The Source of Light

The light that brightens the screens of LCDs has to come from somewhere. There are two lighting methods. First, LEDs at the edges of the screen can use diffusers to light up the entire screen as uniformly as possible and, second, LEDs can be in lines or as a matrix behind the glass plate and again use a diffuser to spread the illumination uniformly over the entire screen area. Smaller screens, like those on cell phones or iPads, use side illumination. Larger screens, like TV sets, use the matrix LED arrangement.

The diffusers are necessary to avoid differences in screen illumination or glare. When a beam of light hits a particle in the diffuser, the light is scattered in all directions. A transparent substrate with loose particles can act as a good light diffuser. A serrated sheet that creates tiny prisms that help scatter the light can also be used. There is usually a reflector in the back of the LED to reflect the light back to the liquid crystal area.

14.3.8 The Entire Operation

Now that we understand all the components of an LCD, lets summarize its operation. Figure 14.12 is the same Figure as 14.5 except that I assume that only the CMOS of the red pixel of a given cell is OFF, and the green and blue CMOS of the same cell are ON. (What I call here a cell, is called also a pixel and the three different colors are called subpixels). Going from left to right, the white light from the LED (white arrows) comes from the left and goes through a diffuser (gray) that makes the back illumination uniform. Then the light goes through the first polarizer (yellow). The light, the electromagnetic wave, now vibrates in only one direction, crosses the glass and the thin-film semiconductor layer (violet), which has been mostly removed (Figure 14.10), crosses the very thin pixelated contact (black), and reaches the liquid crystal (light blue). In the meantime, the CMOS switches in each pixel (violet) have received the information from the receiver and some of them are turned ON and others are OFF. The pixels that are ON force the molecules in

the liquid crystal to align themselves in the direction of the polarized light and therefore they let the polarized light go through the liquid crystal. I assume that the green and blue pixels have the switch ON so the polarized light of these two pixels (green and blue) goes through undisturbed and hits the second polarizer (yellow), which is rotated 90° from the first polarizer. The green and blue light are therefore stopped by the second polarizer. The red light in the upper pixel, which is OFF, is randomized by the liquid crystal, therefore now the red wave vibrates in all directions, passes through the contact and the red filter, and, since the electromagnetic wave is now randomized, passes through the second polarizer and the red beam shines through.

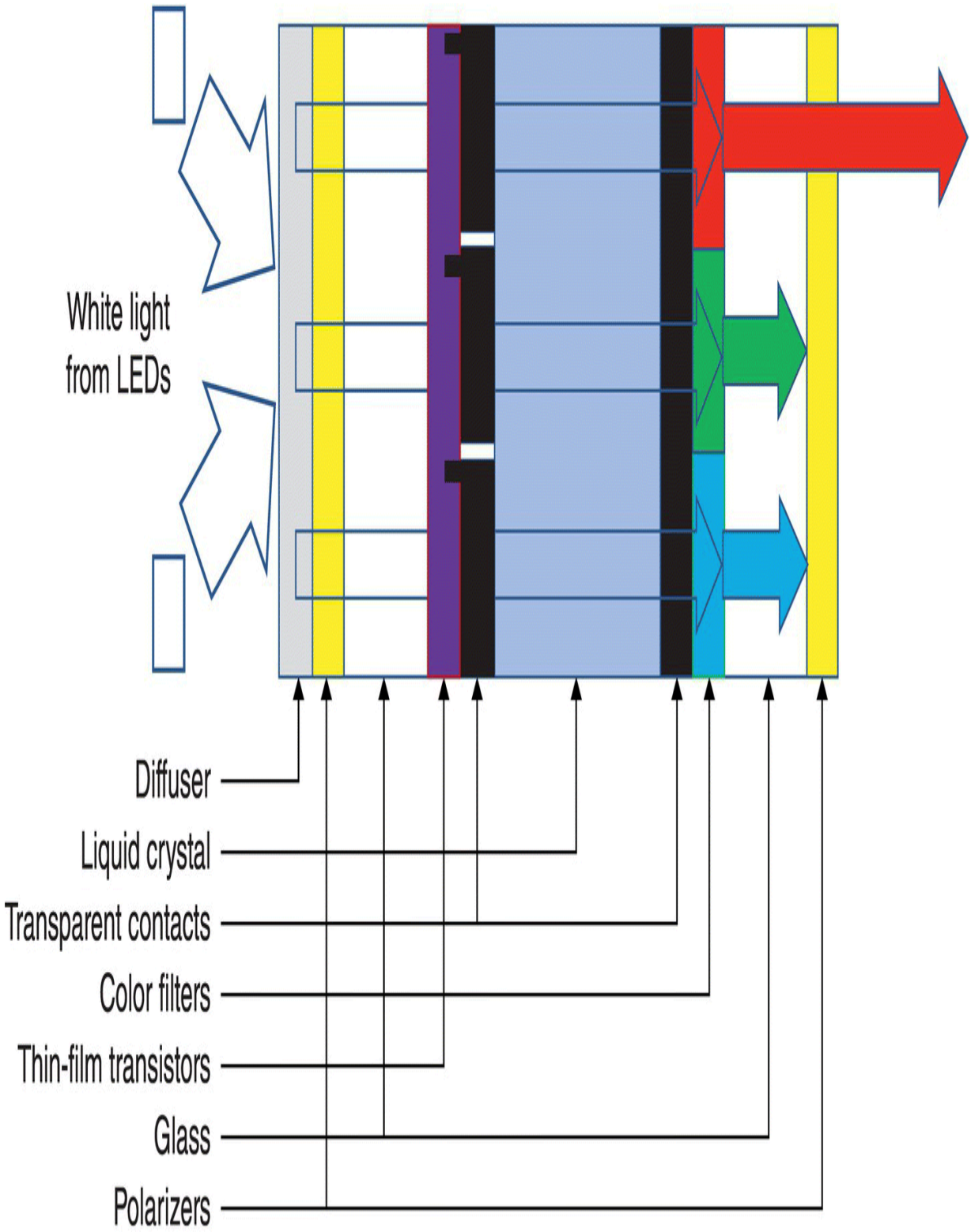

White light
from LEDs
Diffuser
Liquid crystal
Transparent contacts
Color filters
Thin-film transistors
Glass
Polarizers

Figure 14.12 The transistor of the red pixel is OFF, scattering the light inside the liquid crystal and allowing the light to go through the second polarizer. The other two transistor pixels are ON, therefore the molecules are not scrambled and the beam is blocked by the second polarizer.

14.4 Summary and Conclusions

In this chapter we have put together the many components that we have covered in previous chapters, not only taking a look at the computer and microprocessor, but also having fun seeing how the ubiquitous LCD screens work. In the next and final chapter I briefly discuss (speculate about?) where semiconductor technology goes from here.

Appendix 14.1 Keyboard Codes

I mention in the text that one of the reasons the byte contains 8 bits is because 1 byte can represent all the keys of a keyboard. The 8-bit digital number 11111111 is equivalent to 255 in the decimal system. Table 14.1 shows the ASCII code (American Standard Code of Information interchange), which assigns a digital number to each letter, number or special character you find on the keyboard. For example, the capital letter L is assigned the number 76 or the digital number 1001100. The highest number assigned to the keyboard is for the dash (−) and is 126, or 1111110. This last number has 7 bits. We have a bit left in the byte and this first bit is used for the sign of the number. If the first bit is 1 the number is negative, and if it is 0 it is positive. That is a total of 8 bits and why a byte is defined as an 8-bit number.

Table 14.1 The ASCI code.

Letters				**Number**	**Special characters**	
65 A	78 N	97 a	110 n	48 0	32 space	58 :
66 B	79 O	98 b	111 o	49 1	33 !	59 ;
67 C	80 P	99 c	112 p	50 2	34 ”	60 <
68 D	81 Q	100 d	113 q	51 3	35 #	61 =
69 E	82 R	101 e	114 r	52 4	36 $	62 >
70 F	83 S	102 f	115 s	53 5	37%	63 ?
71 G	84 T	103 g	116 t	54 6	38 &	64 @
72 H	85 U	104 h	117 u	55 7	39 ’	91 [
73 I	86 V	105 i	118 v	56 8	40 (	92 \
74 J	87 W	106 j	119 w	57 9	41)	93]
75 K	88 X	107 k	120 x		42 *	94 ^
76 L	89 Y	108 l	121 y		43 +	95 _
77 M	90 Z	109 m	122 z		44 ’	96 `
					45 –	123 {
					46 ·	124 |
					47 /	125 }
						126 ~

15
The Future

OBJECTIVES OF THIS CHAPTER

Predicting the future is always problematic. Someone said that the probability of being right is always 50%: I am either right or wrong, either yes or no, I win or I loose. In spite of that, I will attempt to give you an idea of progress in semiconductor technology and some of the thoughts engineers have about the future.

In the 1960s the first silicon wafers fabricated were about 1 in. or 25 mm in diameter. Now we are already using 300 mm wafers, about a foot in diameter. Considering that the area is proportional to the square of the radius, we are talking about an increase in area of almost 150 times. Can we increase this further? Yes, companies are now working to develop 450 mm wafers. This increases the area by another factor of 2.25. This increase will make the die cheaper but the investment (developing the technology and modifying all the process equipment to accommodate the larger wafer size) is enormous. This, by itself, may be an improvement in the cost of the chips but not in their performance. Furthermore, the larger the wafer the higher the probability that there may be errors and defects, which will make the chips not as cheaper as we may expect.

15.1 The Past

As historians like to say, we can learn about the future by looking at the past. So, let me review what has happened in the past.

1900s: Development of the first vacuum tube, the triode, and the beginning of the electronic age. The feature size (by feature size I mean the size of a basic active component) was about 10 cm.

1940s: Demonstration of the first transistor using germanium.

1950s: Development of the silicon process (pure silicon growth, photolithography, depositions, etc.) and the first silicon transistor BJT. The wafer size was 1.5 in. and the feature size was 1 cm.

1960s: We see the first MOSFET and CMOS. The feature size has decreased by a factor of 10 to 1 mm.

1970s: The surge of microelectronics with 3-in. wafers and microprocessors with about 1000 MOSFETs. The feature size has decreased by a factor of 100 down to 10 μm.

1980s: 4-in. wafers and the VLSI process in full swing. The feature size decrease by another factor of 10 to 1 μm.

2000s: With 12-in. wafers we start the area of nanoelectronics with a feature size of 100 nm.

2018: Now multicore devices use 300 nm wafers and the feature size has decreased to 15 nm.

Since the 1940s we have decreased the size of the components we use by a factor of one million. The size of the devices, measured by the feature size, has also decrease by a factor of a million. The diameter of the wafers we process has increase by a factor of 20 and the area by a factor of 150.

In early 1960 I was studying solid-state physics at Northwestern University. We had an analog computer that occupied a wall about 25 ft long and 8 ft high (Figure 15.1). We had a technician who came to the laboratory early every morning to change all the vacuum tubes that had failed the previous day. We now carry in our pockets microprocessors that are hundreds of times more powerful than that computer.

Gordon Moore (born in 1929; Figure 15.2) the past CEO of Intel, predicted in 1955 that the density of transistors in a chip would double every two years. Was he right? Take a look at a graph of the progress of semiconductor silicon technology (Figure 15.3). The graph shows the exponential growth (linearly in a logarithmic scale) of the number of transistors that we are able to place in a mm^2 area. It was about 200 in 1970 and about 10 million today, which is a growth of 26% per year, exceeding Moore's prediction of 50% every two years.

I show in Figure 15.4 another way of looking at Moore's law. I have added to the graph a trend line from 1970 to 2016 that shows the increase in the number of transistors in an integrated circuit. If I calculate this slope correctly, this growth shows a 37% increase every single year for the past 50 years. This performance improvement is similar but not the same as the previous graph since the chips are not the same size.

The final chart I want to show is in Figure 15.5. This plot looks not at the individual number of devices in a chip or their size, but at the overall performance of the entire chip, including speed, power, and architecture. There are lots of details in Figure 15.5 but I just want to emphasize the growth pattern. In the first eight years, the performance growth was a healthy 25% a year. The phenomenal performance increase in the next 17 years (1986–2003) was a huge 52% per year. Think about it. If I start at 100, a 50% increase is 50 taking me to 150. The next 50% increase is now 75, the next one is 113 and still the next is 169, etc. so each year the increase in performance is higher and higher and that went on for 17 years. The improvement in the last year of this period went from 4195 in 2002 to 6043 in 2003, an increase of 1848 or 44%. But after 2003, the growth started decreasing to 23%, then 12% and finally 3.5% in the last three years. Is this the end of the growth? No, there are no recessions in this technology but there is a probability of some stagnation. In the next section I cover some of the limitations of silicon technology. It looks like we are hitting a wall. We need new

technologies. I will discuss very briefly what engineers and scientists are playing with to continue the growth. The textbooks that I used when I was studying semiconductors in the 1970s, such as Grove's *Physics and Technology of Semiconductor Devices*, which talks about transistors and FETs, was published in 1967, more than 50 years ago and the theories are still the same as the ones I explain in the first few chapters of this book.

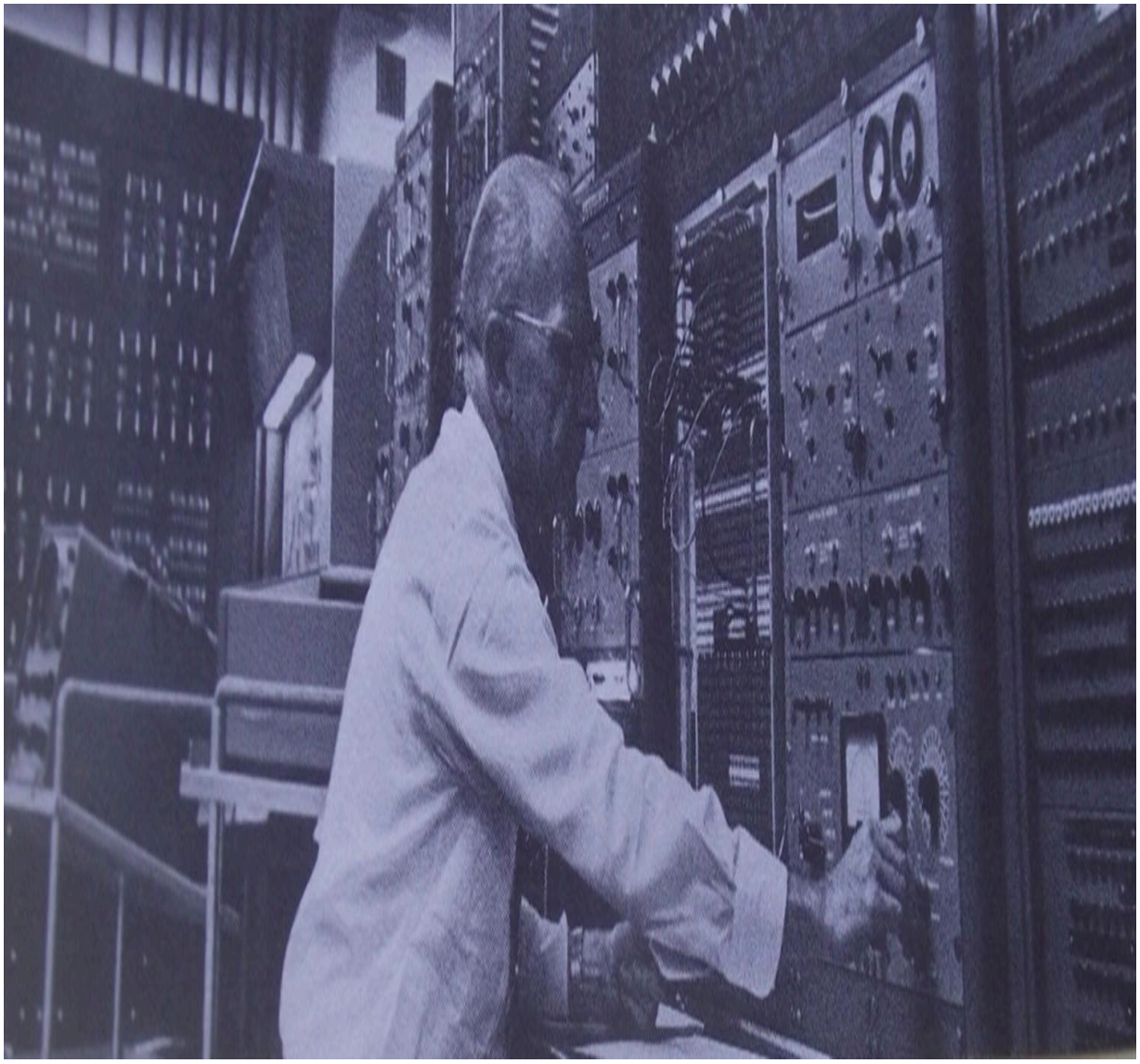


Figure 15.1 Analogue computer at Northwestern University in the 1960s with Dr. Jensen in charge.

Figure 15.2 Dr. Gordon Moore, past CEO of Intel, most famous for Moore's law.

Source: https://en.wikipedia.org/wiki/Gordon_Moore#/media/File:Gordon_Moore_and_Robert_Noyce:at_Intel_in_1970.png.

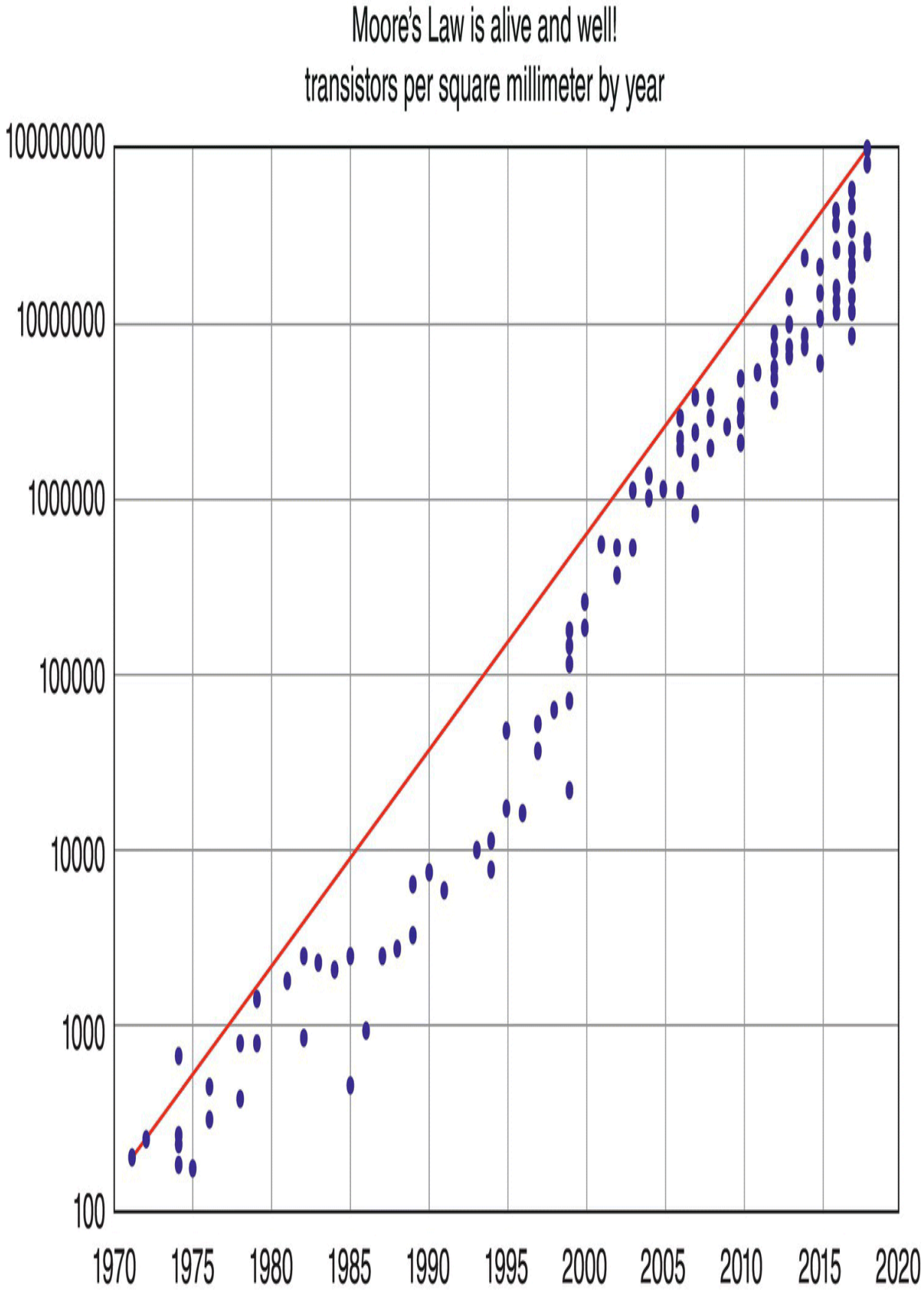

Moore's Law is alive and well!
transistors per square millimeter by year
100000000
10000000
1000000
100000
10000
1000
100
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020

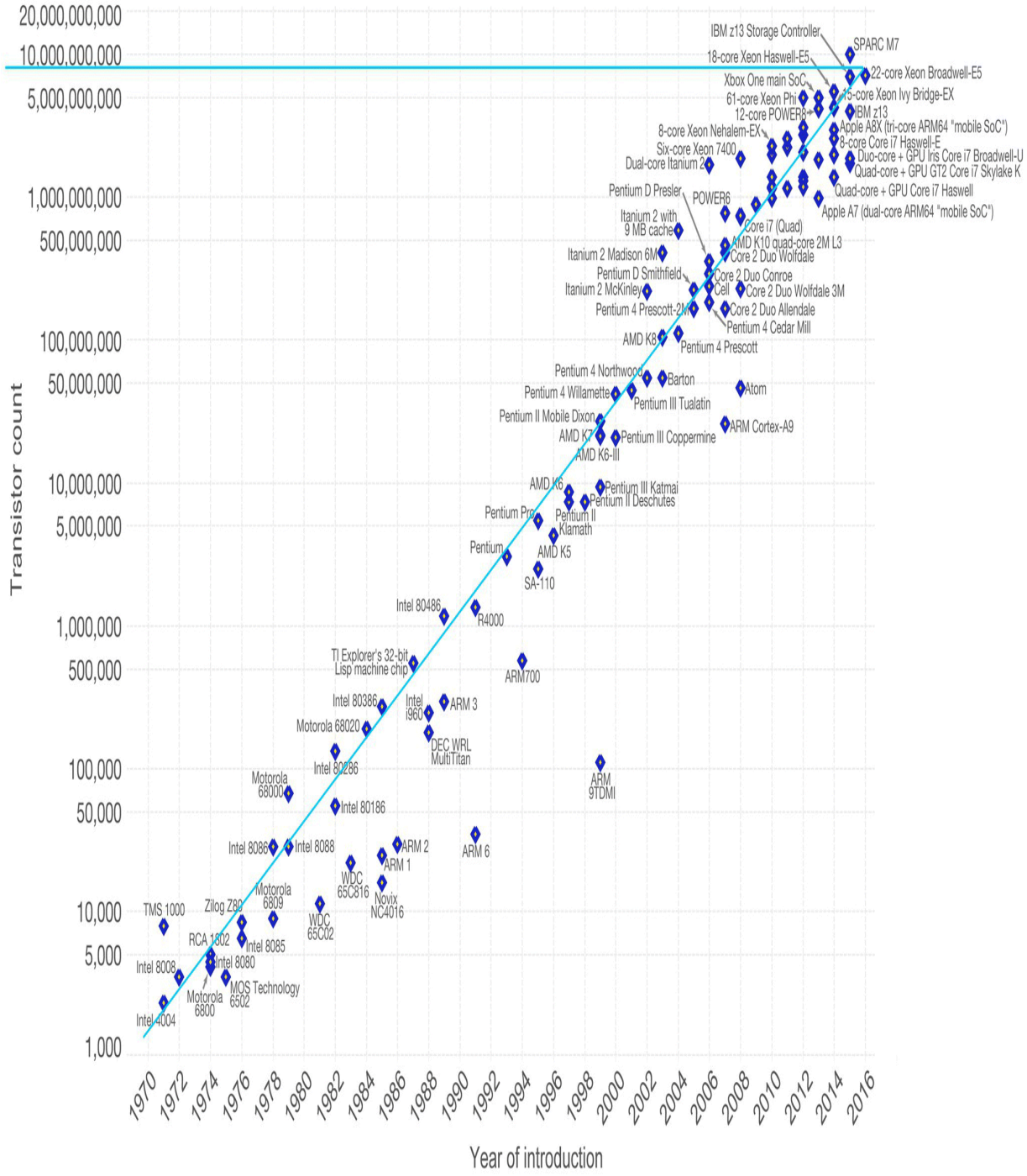
Figure 15.3 The number of transistors in a millimeter square space as a function of time shows a gain of 26% every year since 1970.

Source: https://medium.com/predict/moores-law-is-alive-and-well-eaa49a450188.

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Our World in Data

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)

The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Figure 15.4 The growth of the number of transistors in an integrated chip between 1970 and 2016.

Source: https://en.wikipedia.org/wiki/Moore%27s_law#/media/File:Moore%27s_Law_Transistor_Count_1971-2016.png and https://ourworldindata.org/uploads/2013/05/Transistor-Count-over-time.png.

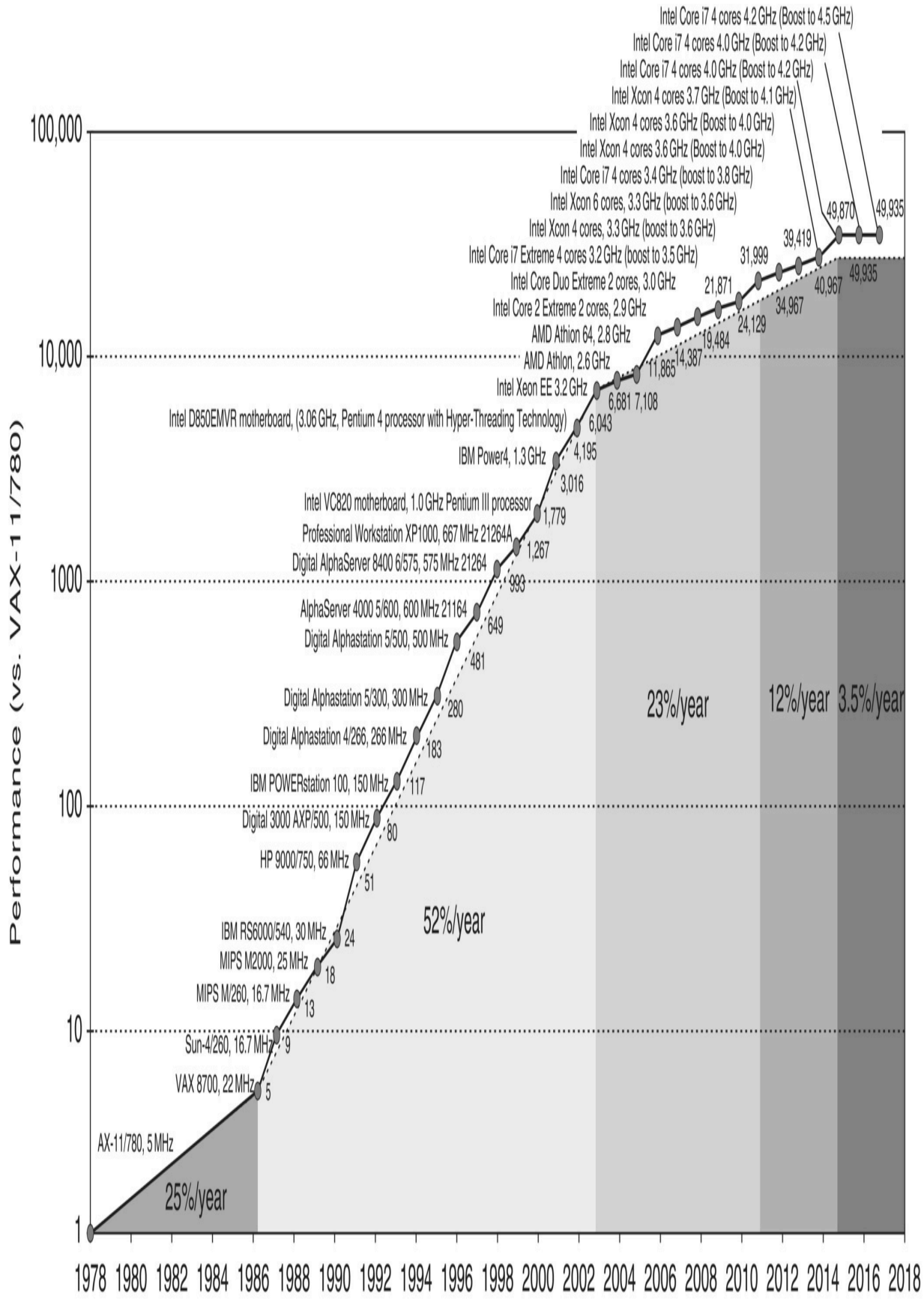

Performance (vs. VAX-11/780)
100,000
10,000
1000
100
10
1
1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018
Intel Core i7 4 cores 4.2 GHz (Boost to 4.5 GHz)
Intel Core i7 4 cores 4.0 GHz (Boost to 4.2 GHz)
Intel Core i7 4 cores 4.0 GHz (Boost to 4.2 GHz)
Intel Xcon 4 cores 3.7 GHz (Boost to 4.1 GHz)
Intel Xcon 4 cores 3.6 GHz (Boost to 4.0 GHz)
Intel Xcon 4 cores 3.6 GHz (Boost to 4.0 GHz)
Intel Core i7 4 cores 3.4 GHz (boost to 3.8 GHz)
Intel Xcon 6 cores, 3.3 GHz (boost to 3.6 GHz)
Intel Xcon 4 cores, 3.3 GHz (boost to 3.6 GHz)
Intel Core i7 Extreme 4 cores 3.2 GHz (boost to 3.5 GHz)
Intel Core Duo Extreme 2 cores, 3.0 GHz
Intel Core 2 Extreme 2 cores, 2.9 GHz
AMD Athion 64, 2.8 GHz
AMD Athlon, 2.6 GHz
Intel Xeon EE 3.2 GHz
Intel D850EMVR motherboard, (3.06 GHz, Pentium 4 processor with Hyper-Threading Technology)
IBM Power4, 1.3 GHz
Intel VC820 motherboard, 1.0 GHz Pentium III processor
Professional Workstation XP1000, 667 MHz 21264A
Digital AlphaServer 8400 6/575, 575 MHz 21264
AlphaServer 4000 5/600, 600 MHz 21164
Digital Alphastation 5/500, 500 MHz
Digital Alphastation 5/300, 300 MHz
Digital Alphastation 4/266, 266 MHz
IBM POWERstation 100, 150 MHz
Digital 3000 AXP/500, 150 MHz
HP 9000/750, 66 MHz
IBM RS6000/540, 30 MHz
MIPS M2000, 25 MHz
MIPS M/260, 16.7 MHz
Sun-4/260, 16.7 MHz
VAX 8700, 22 MHz
AX-11/780, 5 MHz
49,870
49,935
39,419
31,999
21,871
40,967
49,935
34,967
24,129
19,484
14,387
11,865
6,681
7,108
6,043
4,195
3,016
1,779
1,267
993
649
481
280
183
117
80
51
24
18
13
9
5
25%/year
52%/year
23%/year
12%/year
3.5%/year

Figure 15.5 Processor growth in the last 40 years

Source: Computer Architecture by Hennessy and Paterson. Morgan Kaufmann, 2019 (http://www.cs.columbia.edu/~sedwards/classes/2012/3827-spring/advanced-arch-2011.pdf).

15.2 Problems with Silicon-based Technology

To understand the problems and limits that we are reaching with the silicon technology take a look a Figure 15.6. The sketch of the MOSFET is the same the one we have seen before except that I have added more metal layers.

First let's consider what happens to the operation of the component as we make the dimensions shorter and smaller. When we mention that the feature size is 12 nm, for example, we are talking about the distance between the source and the drain of the MOSFET, which we call the *CMOS length* (Figure 15.6). The 12-nm separation means that we have just about 25 silicon atoms between the source and the drain. Now we leave the range of solids and enter the area of atoms. As the distance between electronic elements gets smaller, the thickness has to decrease as well.

Some of the operational problems that need to be solved are the following:

Leakage currents between the gate and the channel as the oxide becomes thinner. The 1-nm oxide layer, is just the thickness of two atomic layers.

Punch-through between the source and the drain is a very serious limitation. You need very little voltage to increase the number of electrons under the gate and the depletion region in the drain, and very soon there is a short between the drain and the source. This punch-through forces us to decrease the drain voltage and the thickness of the oxide or to look for better insulating materials. One possible approach is to grow a very thin epitaxial layer on top of an oxide layer, so the channel is constrained between two oxides.

Tunneling is also a very serious problem as the "barriers" get thinner. No matter how high an electronic barrier is, if it is thin enough the electrons will tunnel to the other side. When the channel is supposed to be OFF, electrons can tunnel through from source to drain. Tunneling may be a fundamental limit to design rules of the order of maybe 3 nm. Even if many of the dimensions keep on decreasing, the channel length limit will be something around 6 nm.

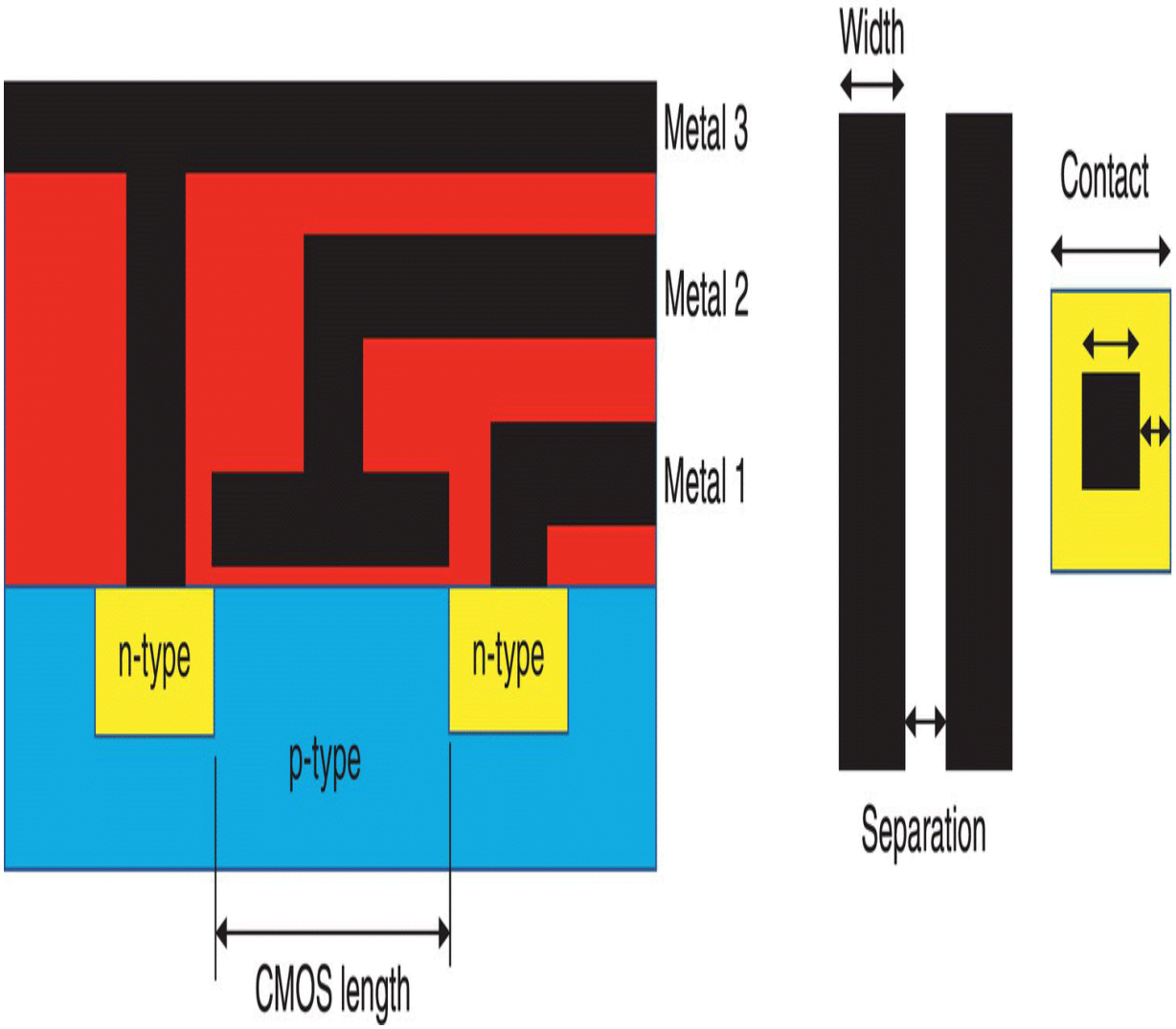


Figure 15.6 A FET and some designs rules that are needed to ensure that key components do not interfere with or touch each other.

Deterioration of silicon properties.

Mobility. As the thickness of the oxide and silicon layers decreases, the mobility also decreases as much as a factor of two as we go from 10 to 4 nm.

Resistivity. As the metal lines decrease in thinness and size, the resistance increases. At these very small dimensions, other mechanisms also decrease the resistivity by as much as a factor of two, making things still worse.

Fringing. The electric field at the edges of a capacitor are not only vertical but, like an ice-cream bar, the electric field bulges at the edges. As the capacitors get smaller and the oxide between the plates thinner, the effect of fringing fields becomes more pronounced.

Power consumption increases as the density of electronic devices and the speed of the microprocessor keep on increasing.

Process variations. It is very hard to have perfectly uniform parameters. Concentration of impurities, oxide thickness, depositions, mask features dimensions, etc. all have some variations. As the feature sizes get smaller, these process variations become much more significant.

Costs increases as the density and number of devices in a chip increase. Most of the costs are due to the initial investment costs of process equipment and facilities. The mask sets are also increasing in cost. These very delicate, complex, and costly masks need to be changed about every 100 uses. The manufacturing costs also increase as the number of layers and the density of components in a chip increase.

The design rules are the way that designers communicate with the foundries that have to fabricate these devices. These design rules prevent shorts between two elements or open contacts. All elements have to be "perfectly" aligned, with the obvious problem that nothing is really "perfect." The fabrication processes have limitations in both accuracy and repeatability.

Design rules include the minimum thickness of the connecting lines, the minimum area of a contact, how they need to be separated, the thickness of the different oxidations, diffusions, polysilicon and metal layers, the width and separation of the metal layers, and how small the vias that connect one to another are supposed to be. These design rules, when we were working with micrometer feature sizes, used to be scalable, that is, going from 150 to 30 μm feature size used to necessitate making (almost) all the design rules five times smaller. These were known as *lambda* design rules. The foundries accepted these numbers and just scaled them for their different lines. As we go to nanometer dimensions, the design rules are no longer scalable; some rules cannot be made as small as the process would indicate. Industry now uses *micron rules* specifying each rule independently and in most cases based on the intuition and experience of the engineers and material scientists.

There are so many possible variations and nonuniformities that it is amazing that these foundries can fabricate the chips successfully. There are variations in the thicknesses of metals and oxides which affect the capacitance and resistance of the devices, variation in impurity concentrations and changes in dimensions due to optical photolithography processes. Implants and diffusion also can cause problems. All of these problems are magnified as we try to get smaller and smaller dimensions.

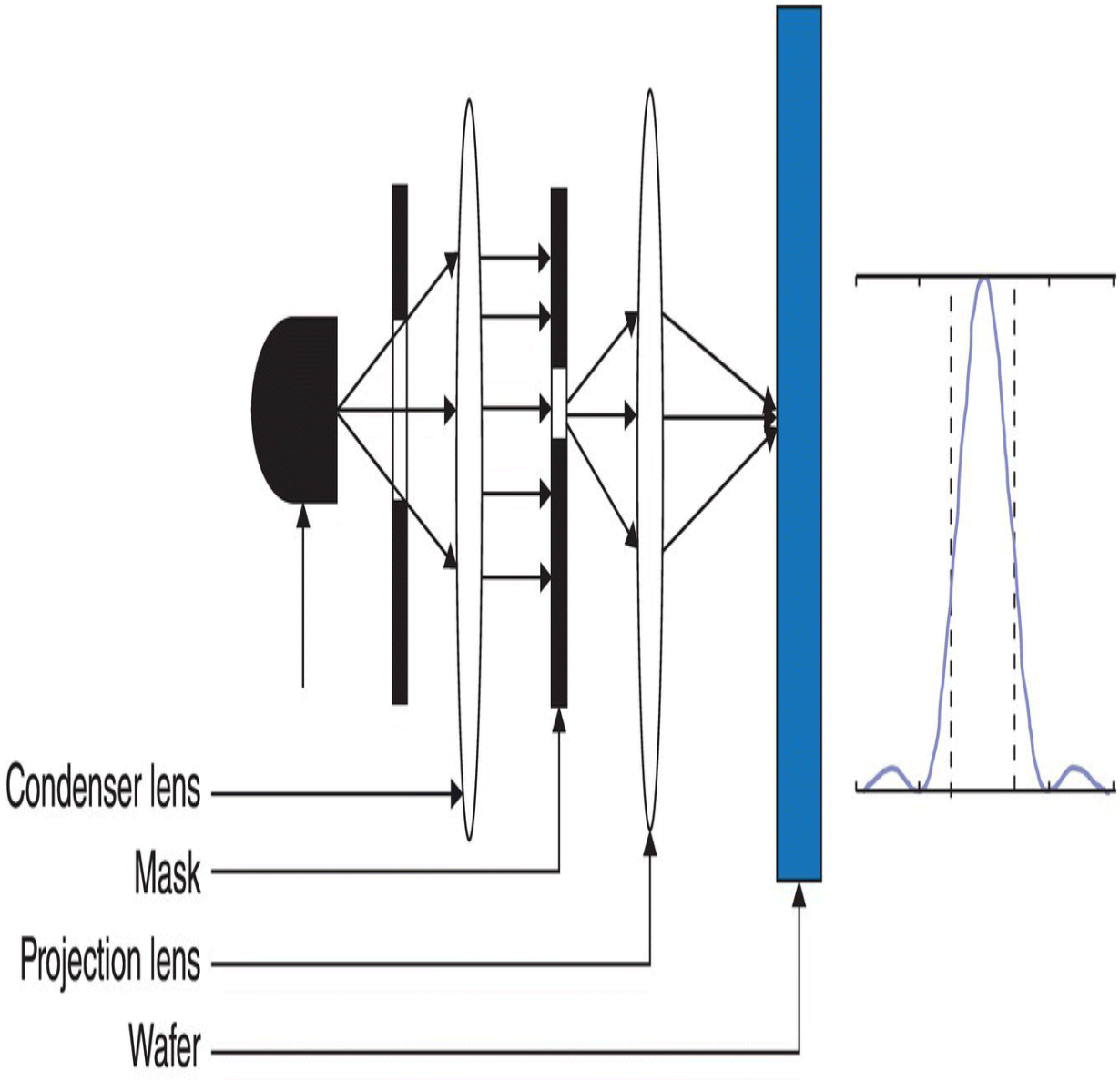


Figure 15.7 Sketch of an optical projection system (left) and the resulting image (right).

Another big problem as we reach smaller feature sizes is the optics. There are different ways of processing a die or a wafer. One is contact printing, in which the mask sits on top of the wafer. The resolution is less than 0.5 μm and the mask is easily damaged. Projection printing is better, down to 0.2 μm using ultraviolet light. I show a typical simplified optical projection system and the resulting image in Figure 15.7. The resolution and the depth of focus is proportional to the wavelength. Therefore, the resulting image is not

the step function we desire (the dotted lines at the right of Figure 15.7), but due to the diffraction effects distorts the square pulse and the distortion gets worse as the wavelength of the source gets longer.

We cannot use visible light as the source for photolithography. The visible light wavelength, as we have seen, goes from 700 to 500 nm, 50 times larger than the features sizes we want to image. We use ultraviolet light that has a range from 400 nm for the near-ultraviolet to 10 nm for the extra-ultraviolet. As we look to go down to 10 nm or even 5 nm features, we have to start looking at using X-rays that go all the way down to 0.1 nm, but we have to find sources, photoresists (i.e. materials that become transparent or opaque to X-ray radiation), and masks (devices that let X-rays go or not go through them (e.g. gold, tantalum or tungsten). The resists must have high sensitivity to X-rays as well as resolution resistance to etching.

Finally, we have to realize that these photolithographic processes are repeated 25–50 times on the wafer and it may take up to three months to finish a lot. Any error in any one of these processing steps may be irreparable and catastrophic.

Because of the lower wavelength, less than 1 Å, X-rays result in sharper images and lower diffraction effects, and the longer depth of focus compensates for the wafer planarity. Ion beam photolithography is sharper than X-rays, but it has a very low penetration which means the photoresist has to be very thin.

Now that I have scared the heck out of every reader and made integrated circuit fabrication almost impossible, let me be more positive. In the next section I discuss some of the technologies outside semiconductor technology that engineers and scientists are working on to bypass the limitations of silicon technology and in the final section we'll take a look at actual research to increase the capabilities of the old reliable semiconductor and silicon technologies.

15.3 New Technologies

Before I talk about silicon technologies that can continue to improve, I like to briefly go over some of the different technologies that scientist and engineers are looking at to replace and improve the performance of computing.

15.3.1 Nanotubes

Nanotubes are quite an interesting material. They are basically carbon in the graphene state. Carbon has four valence electrons that can form two crystallographic structures, the diamond structure in silicon (Figure 3.1) and a second graphite structure (Figure 15.8A).

Covalent bonding is one of the strongest and diamond is one of the hardest materials. The other crystallographic structure of carbon is graphite, which consists of hexagonal planes using three of the valence electrons to join hands with the other carbon atoms in the same plane (A) and the fourth electron connecting with a corresponding electron in an adjacent plane (B). The horizontal lattice distance between two atoms on the plane (A) is 0.14 nm, very strong covalent bonding, but the vertical distance between the atoms in different planes (B) is 0.34 nm, about 2.5 times longer. Thus, the strength between planes is very weak. Graphite sheets can be easily separated from one and another. As you intuitively know, graphite is the most common carbon crystallographic structure in nature and thus is considerably more abundant than diamonds.

The interesting thing about these graphite sheets is that they can be wrapped around to form a tube (Figure 15.8C). The ratio between the diameter and the length of the nanotube can be as large as 1:1000.

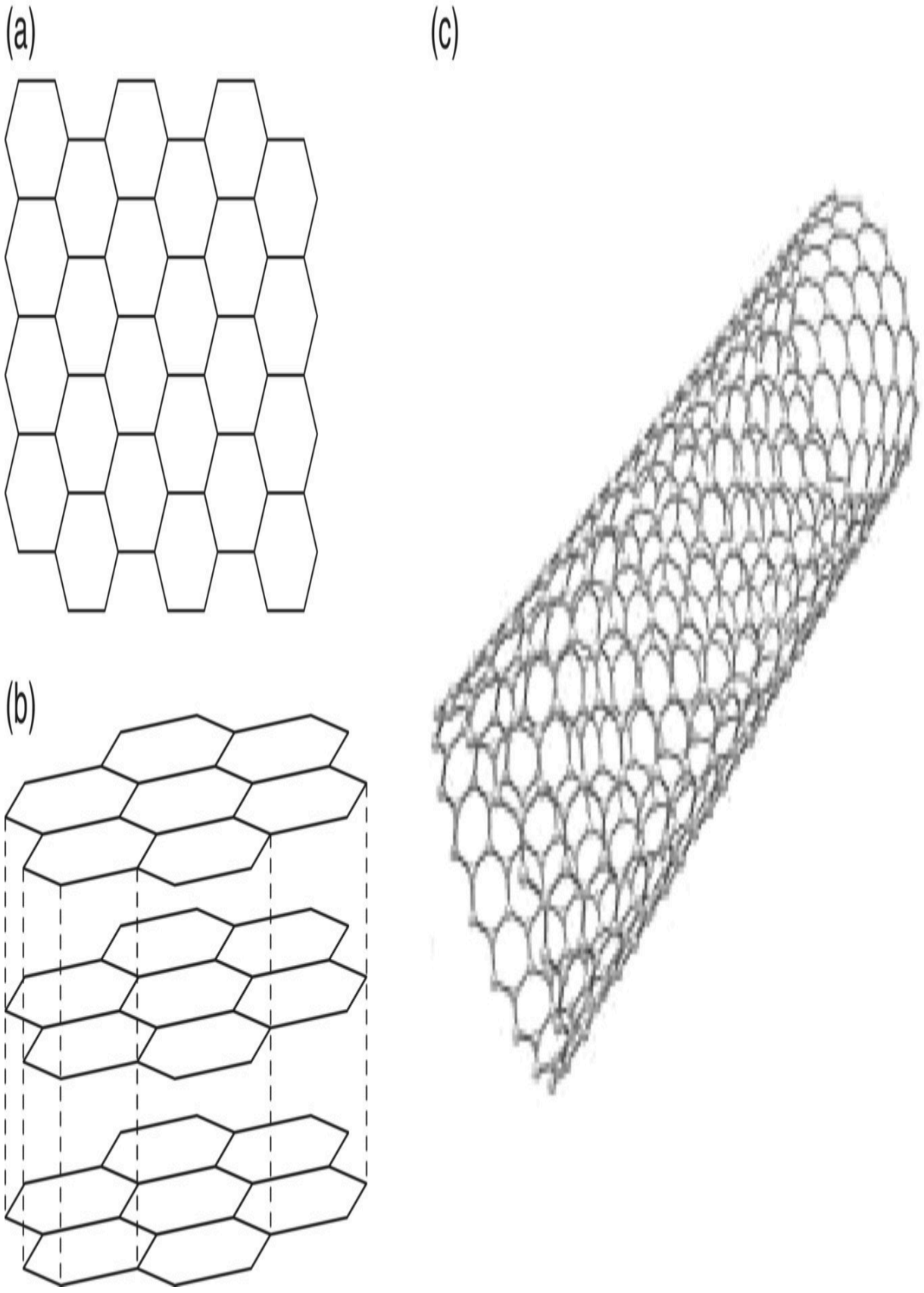

(a)
(b)
(c)

Figure 15.8 Crystallographic structures of carbon in the graphite state (A and B) and forming a nanotube (C).

Source: https://www.123rf.com/stockphoto/nanotubes.html?&sti=o17d96v5t87ilqwivp|&mediapopup=92428614.

Carbon nanotube transistors are expected to be five times faster than silicon transistors. They are also very strong because they use strong covalent bonding.

The free electrons move along the tube. The energy levels of the nanotube can be calculated by analyzing the two-dimensional crystal or the graphene. The bandgap depends on the diameter of the nanotube. A nanotube can replace the channel of a MOSFET. It will be much smaller and much faster. These nanotube devices have already been demonstrated in the laboratory. The problem now is to make them manufacturable.

In November 2019, the *New York Times* published an article reporting quite a development in nanotubes: https://www.nytimes.com/2019/10/30/science/graphene-physics-superconductor.html?searchResultPosition=1. Researchers had been able to make a single graphene sheet, one atom thick (by the old trick of peeling off layers with sticky tape). It is a very interesting article showing the promise of this technology.

15.3.2 Quantum Computing

There are three concepts needed to understand quantum computing.

Qubit: This is an electron or a photon or any other particle that can have two quantum states, like the two spins of electrons or the horizontal and vertical polarization of photons.

Superposition: Like the classical bit, the qubit can have either state 1 or state 0, but it can also have a combination of 1 and 0 by superposition. Its value, 0 or 1, is unknown until we measure it. The output of a measurement will always be either a 1 or a 0. It is the

case of a cat inside a box which is both dead and alive at the same time, until we open the box and look inside, or do a measurement.

Entanglement: If two qubits are "entangled" they can be separated by very large distances, and knowing the properties of one we can immediately know the properties of the other. If I change the properties of one, instantaneously the properties of other change. This is what is called *coherence*. So, if I measure one qubit, I know the state of all the qubits entangled with it.

If you have difficulty accepting these concepts you are not alone. Einstein initially had lots of problems with these concepts. He thought that the cat dead or alive made no sense, "God does not play dice" he said, and entanglement was action at a distance, moving faster than the speed of light, a kind of "quantum teleportation."

A number of N classical bits can uniquely identify 2^N numbers. For example, if the number of bits is five, $N = 5$, we can identify 32 numbers, from 00000 to 11111 (from 0 to 31). The important point is that at all times we know which number out of the 32 possible ones actually exists, it is, for example, 10101 and no other. In quantum computing all numbers are there, not just one. When we measure it, out of all the 32 numbers, the result will give you the most probable result.

In quantum computing the first thing to do is to prepare the qubit quantum state, then perform a quantum operation on the qubit, and finally read the output, which is by nature probabilistic. This is repeated several times until we are happy with the probabilistic result. The more times we do the operation, the surer we are of the result.

Because the result is probabilistic, there is always the possibility of an error so you may need to check again, but it will still be much faster than if you were to do the operation on the fastest classical computer that we have right now.

Suppose you want to find the largest number in a database that has 10^8 entries. The classical computer has to look at all the entries one at a time. The quantum computer makes this selection by the square root of the number of entries, or 10^4. So even if you want to repeat the calculation 10 or more times to be sure that the probabilistic result is correct, it is still 1000 times faster than the classical computer.

Suppose you have five students in a class, and you want to know how many have the same birthday. You would pick the birthday of student one and compare it to the other four, then student two and compare their birthday to the other three. This process requires 10 checks. But if you had 1000 students, you would need to do the comparison 500 500 times. You can see that these types of calculations grow exponentially as the number of items grows. That is one of the reasons why quantum computing is so appealing.

Although a quantum computer is not one you would have on your desk, at least not in the foreseeable future, it still will help perform calculations that are now too complex, for example in medicine or chemistry, looking at the interaction of molecules. Right now, the modeling of a simple molecule, looking at all the electron interactions, can take months.

One of the problems of quantum computing is *decoherence*, that is, the loss of coherence. Information gets corrupted because it interferes with the surrounding environment. Coherence can last for 100–200 μs, depending on how well the system can be isolated from the environment.


IBM Research
Quantum Computing
73

Figure 15.9 An IBM quantum computer.

Source: https://www.shutterstock.com/image-photo/hannover-germany-june-13-2018-ibm-1157800963?src=map7dfvYWyIxKzK0ubTgPw-1-0www.research.ibm.com/ibm-q.

In October 2019 the *New York Times* published two articles announcing a breakthrough in quantum computing. Google, using quantum computing, calculated in 3 minutes and 20 seconds what would have taken 10 000 years to calculate using a standard computer. You can read the two articles here: https://www.nytimes.com/2019/10/21/science/quantum-computer-physics-qubits.html and https://www.nytimes.com/2019/10/23/technology/quantum-computing-google.html. At that time, and the time of this writing, these results need confirmation, but this shows the capabilities of the quantum computing. These two articles, by the way, explain quantum computing technology quite well.

It now seems inconceivable that we will have a quantum computer on our desk. The IBM quantum computer (Figure 15.9) occupies a whole laboratory and has supporting electronics and liquid helium tanks to cool the devices to close to 1 K, but if you go back to the 1960s (Figure 15.1), we would never have thought that a computer that then occupied an entire room could possibly be carried on a briefcase.

15.3.3 Biocomputing

It happens that some organic molecules have properties similar to semiconductors, with mobilities and energy gaps. In addition to being able to deposit very thin layers of these organic components, they are very flexible. Imagine rolling up a computer and carrying it in your shirt pocket. The energy levels are different for different polymers and can be doped by adding impurities that capture an electron and therefore it leaves a hole and both can move around. Sounds familiar? This charge in an organic semiconductor has been named a *polaron*. The impurities change the conductivities of

organic films. The energy level differences are appropriate for interaction with visible and infrared radiation. The main molecules used are DNA and proteins.

It is interesting that some of the fabrication methods are quite similar to those used in semiconductor processes. We can deposit metals on top of the organic films. These are used as masks. The films can be made as small as 2 nm. These devices will be very fast and dissipate very little power.

One thought-provoking concepts is the fact that all biological molecules have the capability for self-replication, that is, they have the ability to grow. I mentioned before the possibility of rolling up your computer, now we also see the ultimate possibility that the computer will grow and self-reproduce. You buy a biocomputer and with proper tender care you can have several computers for the entire family. But now I am in the realm of science fiction. Jules Verne in the late 1800s predicted many of the technologies that we have today.

15.4 Silicon Technology Innovations

Let's finish the book with something more down to earth, at least for the first half of the twenty-first century. At this time there are no short-term candidates to replace the old, reliable, proven, CMOS technology. Silicon technology will be with us for many more years. Can we improve the technology? Yes, in spite of all the difficulties I mentioned in the first section. Here are a few possible improvements.

15.4.1 Process Improvements

I have already mentioned the research and development that has been done to fabricate 450 mm wafers (some are already toying with the idea of going to 675 mm wafers) which will increase the area by more than a factor of two and thus allow twice as many devices or the same devices twice as large to be fabricated.

The next improvement is in the photolithography so we can get down to the 7 or 5 nm feature size. If successful, this will quadrupole the number of components that one can place on a chip. This will not only increase the component density but also, because components are closer, increase the speed and allow many more parallel computations. Actually, there are masks with varying thicknesses (called phase shift masks) so that the shift cancels the light where we do not want it, which improves the resolution.

The semiconductor community is also investigating new materials. New oxides with higher dielectric value can be deposited and they are thinner and harder than the SiO_2 we use today. Also, metals and doped polysilicon decrease the resistance of the lines. As we get to smaller feature sizes copper is taking the place of aluminum.

Another technology that is being implemented is silicon-on-insulators (SOI). Figure 15.10 shows an implementation of this technology. The main advantage is that the devices, n-MOS and p-MOS, are totally isolated. This technology allows a higher component density and lower parasitic capacitance, but (there is always a "but") it is a more costly process.

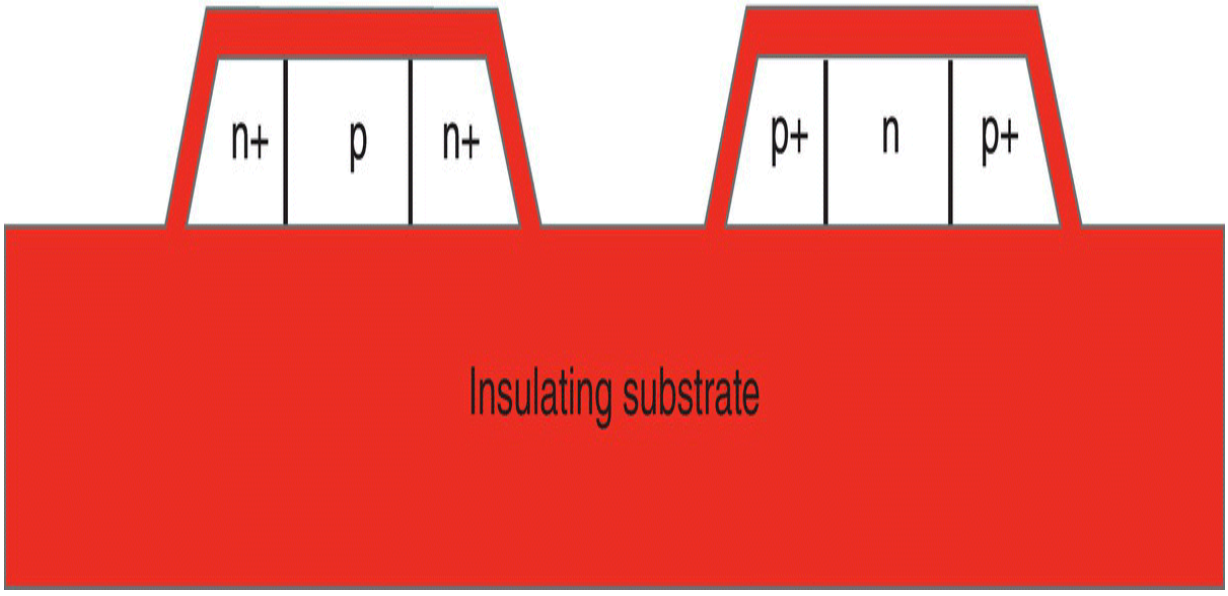


Figure 15.10 An n-MOS and p-MOS fabricated on top of an insulating substrate, completely isolated on all sides.

15.4.2 Vertical Integration

If you think about what we have talked about up to now, silicon technology has been very much a planar two-dimensional technology. But what if we could grow one layer over another? Figure 15.11 shows one implementation of vertical integration. There are two active layers identified a s T1 and T2 and six metal layers, M1 to M6. The first T1 layer could be, for example, the microprocessor and the T2 layer could be one or more of the memories. This type of vertical integration not only increases the component density on the chip, but it potentially increases the data transfer speed dramatically. It is like going one floor up to borrow a cup of sugar versus taking out the car and going to the supermarket.

This vertical integration is already being done both with metal busses and flip bonding. We have seen flip bonding in Chapter 10 (Figure 10.29). The same technology can be used in other applications. One problem is to make the indium bump interconnect sufficiently small and reliable so that we do not use a lot of area for pads or an imperfect contact disconnects a key junction.

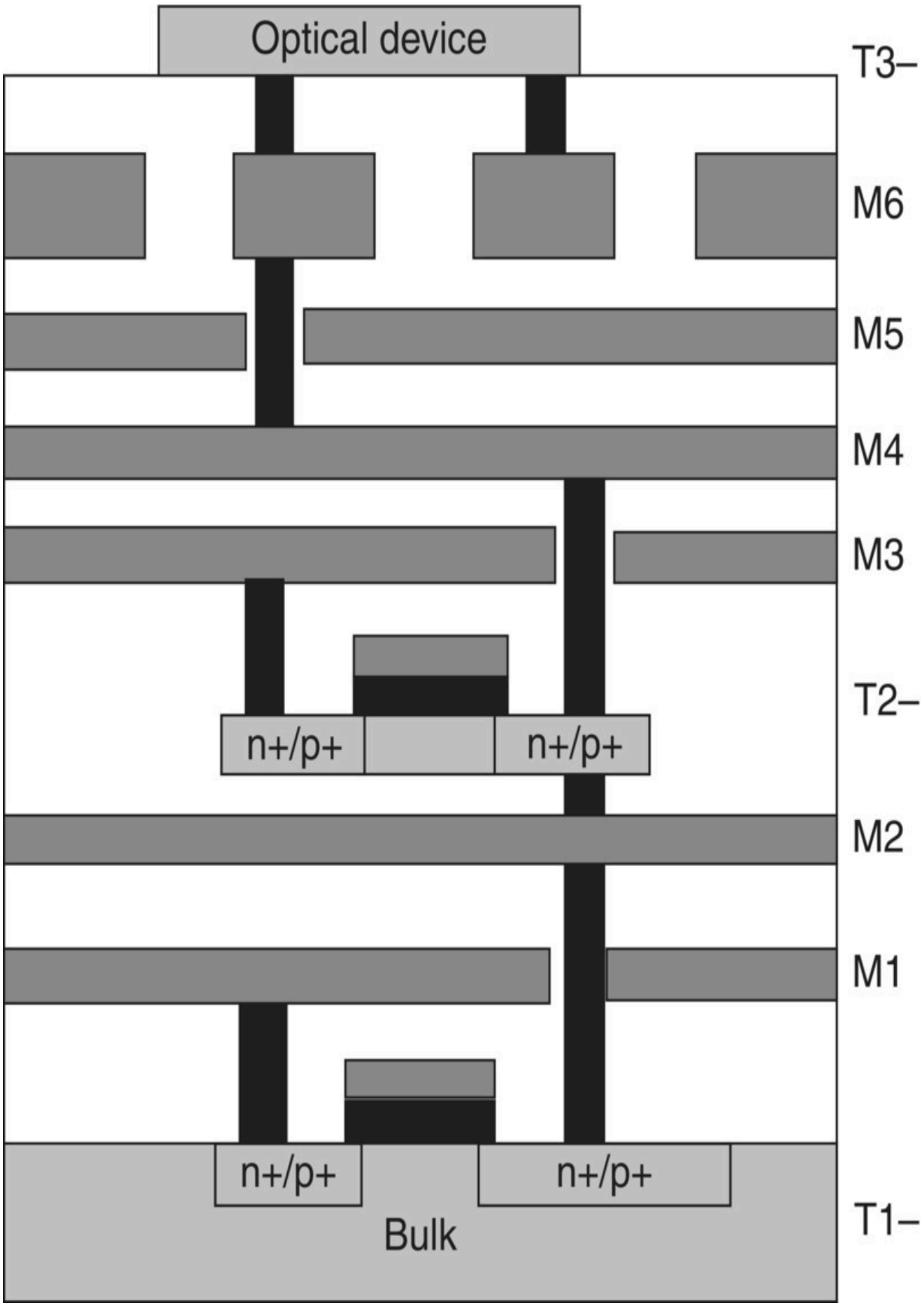

Optical device
T3–
M6
M5
M4
M3
T2–
n+/p+
n+/p+
M2
M1
n+/p+
n+/p+
Bulk
T1–

Figure 15.11 In a vertical integration process we deposit more than one layer on top of another, doubling the circuitry that can occupy the same place.

Source: bwrcs.eecs.berkeley.edu/Classes/icdesign/ee141_f01/Notes/chapter2.

Multiple metal interconnects are already being made, as shown in Figure 15.12. Note that the aluminum lines get thicker as they move up from the electronic components. They are the lines that go farther away so the larger thickness improves the resistance of the line, like a country road versus a four-lane highway. Furthermore, there is more space in the upper layers to lay longer and thicker lines.

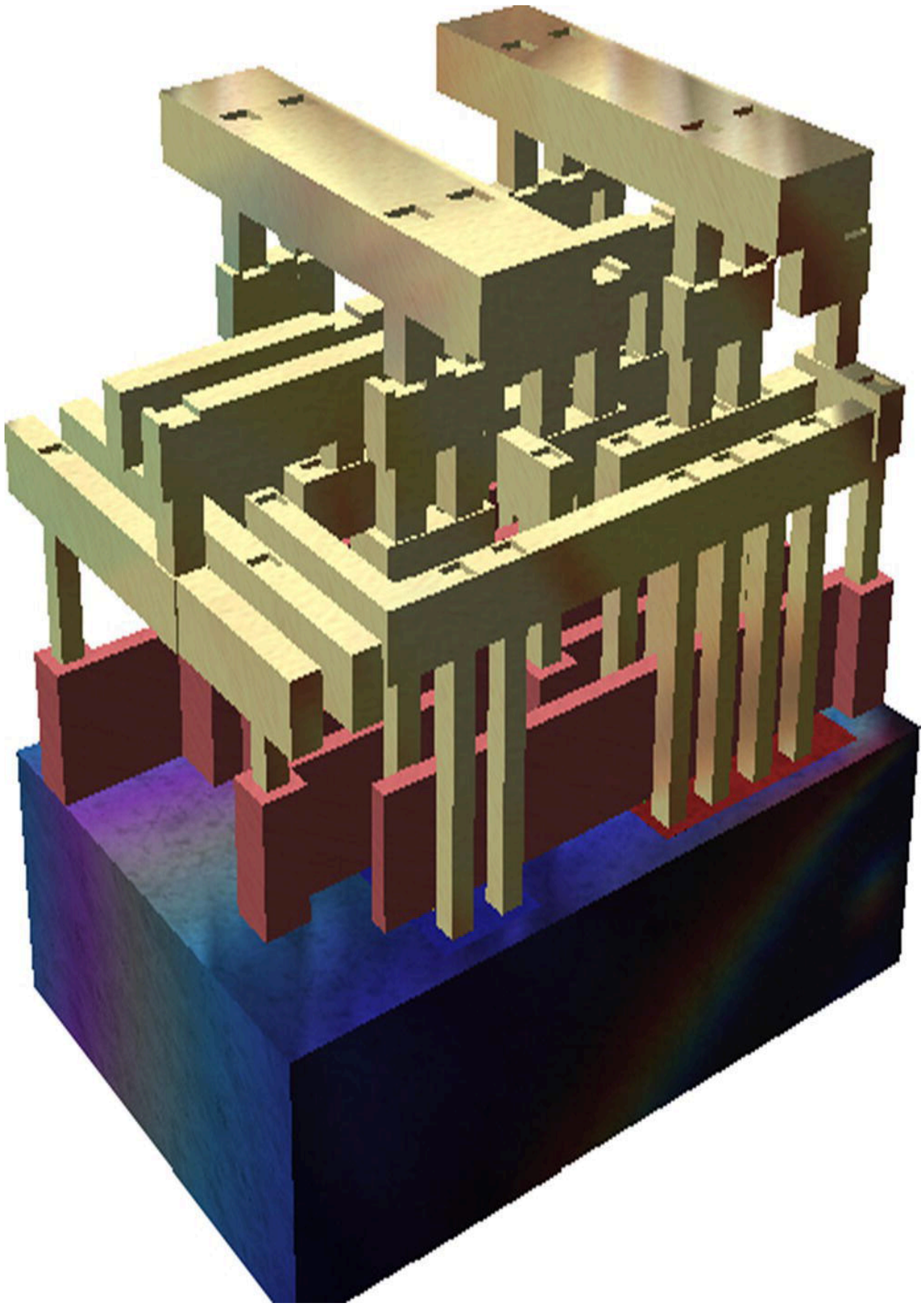


Figure 15.12 An example of multiple metallic layer interconnects.

Source: https://en.wikipedia.org/wiki/Integrated_circuit#/media/File:Silicon_chip_3d.png.

15.4.3 The FinFET

What if we were able to control the gates from all different sides? We have seen that the tunneling starts when the thickness of the barriers goes under 5 nm. We can avoid this law of nature by going up instead of sideways. That is what FinFET does (Figure 15.13) (FinFET is so-called because the FET active components, source, channel, and drains, look like fins on top of the silicon.) The area of the semiconductor channel has increased quite a bit, although its density has decreased. This process allows higher integration density. The current flows parallel to the two plates of the semiconductor. The gates can continue on top, completely surrounding the silicon wall. The small space can carry more current that the planar FET. It is very compatible with the current processes. This technology has less defect density than the planar one. In this technology the gates are wider than the Fin body.

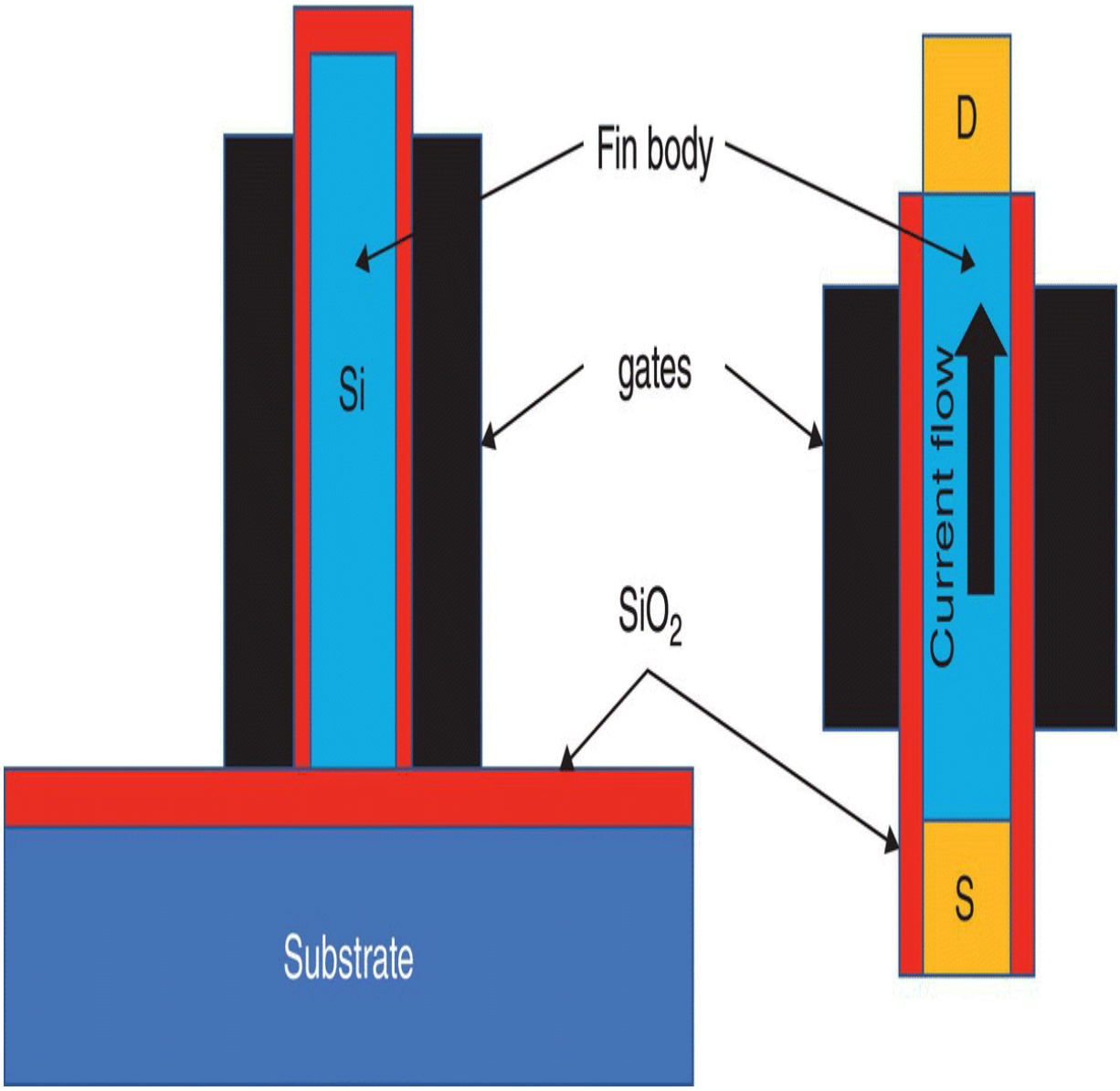


Figure 15.13 Sketch of a FinFET. The semiconductor is a very thin vertical silicon with two gates on either side. I show the side view on the left and the top view on the right.

15.4.4 The Tunnel FET

The Tunnel FET or TFET has a structure similar to the PIN diode I discussed in Section 13.2 except that now the two contacts are replaced by two highly doped p+ and n+ regions, and we add a gate on top of the intrinsic region. The intrinsic region of the PIN diode is

long so that there is a large region to collect as many photons as we can. The TFET has a very thin intrinsic region (Figure 15.14).

In the upper sketch of Figure 15.14 I show the structure of the TFET. It consists of two very heavily doped regions, the source with p++ doping and the drain with n++ doping, separated by an intrinsic region, i. On the lower left I show the band diagram of the TFET when it is reversed biased but there is no voltage applied to the gate. Like in the tunnel diode under reversed bias (Section 5.4) the valence band of the p++ source has higher energy than the conduction band of the n++ drain. There is no current between the source and the drain because the intrinsic region in the middle has no free electrons in the conduction band or free holes in the valence band, and the source and drain are separated by the thick barrier created by the intrinsic region. When we apply a positive voltage to the gate (lower right part of Figure 5.14), we bring the potential in the intrinsic region way down, and now the transition region is very thin and electrons from the p++ valence band can tunnel through to the empty energy sites of the intrinsic and n++ regions.

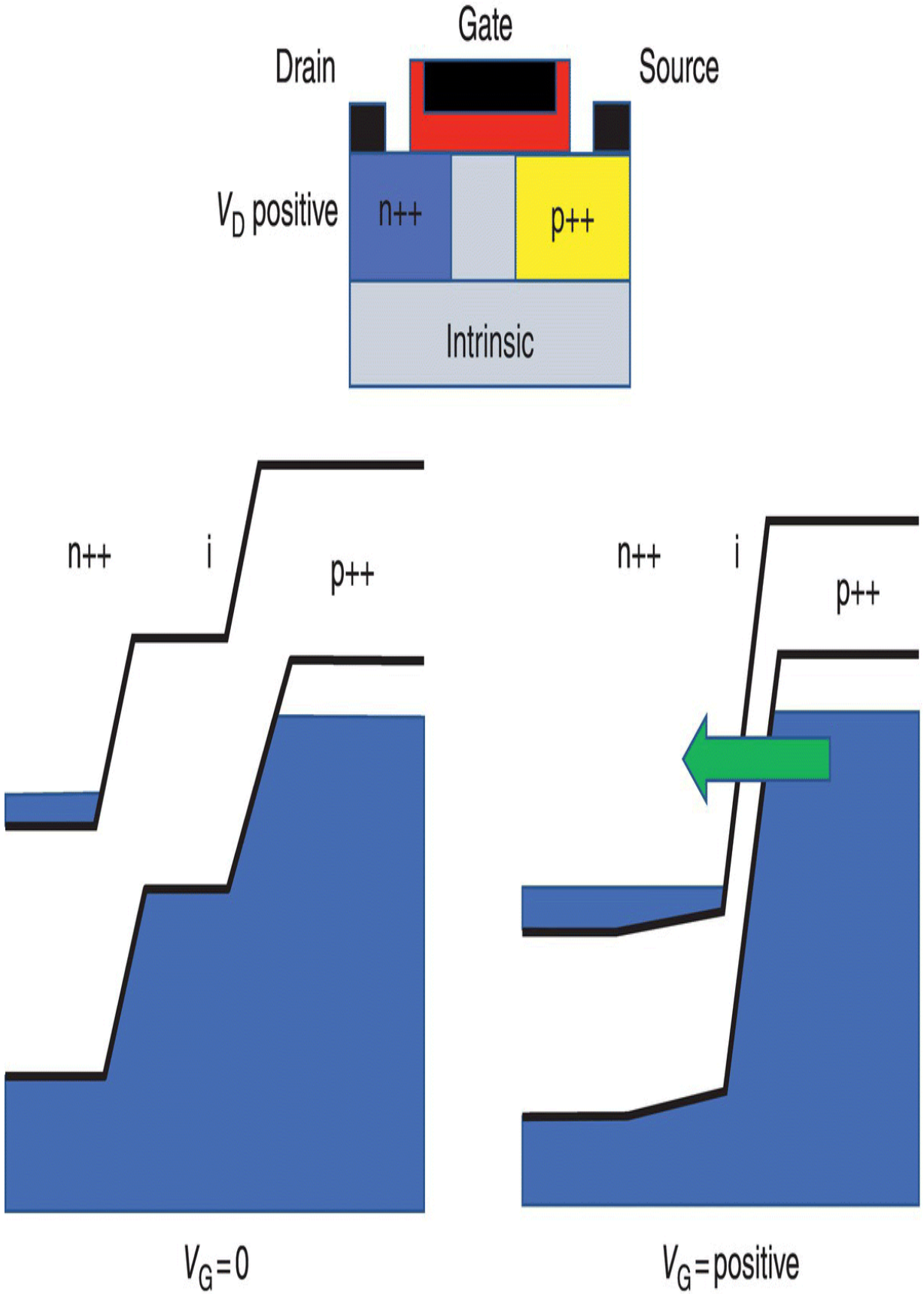

Gate
Drain
Source
V_D positive
n++
p++
Intrinsic
n++
i
p++
n++
i
p++
V_G=0
V_G=positive

Figure 15.14 The tunnel FET and energy bands when the TFET is reversed biased and the gate voltage is OFF (left) and ON (right).

This transition, from ON to OFF, can be done at much lower voltage levels than the regular CMOS transistors and thus reduce the power consumption. Additionally, the TFETs can switch very fast. One problem is that the currents are rather small, so engineers and researchers try to combine the TFET with technologies similar to the three-dimensional FinFET to increase the current without increasing the potential barrier that allows the tunneling.

15.5 Summary and Conclusions

Semiconductor devices have revolutionized electronics in the past 80 years and all indications are that they will continue to be the electronics technology driver for the next 50 years. What a record! Silicon, which is nothing but sand, will still dominate the field. All the technologies I have sketched in this final chapter have been demonstrated in the laboratory and some are already being implemented in some small manufacturing environments. Engineers and scientists will continue to invent new devices and improve others. It will be a treat to see how the technology grows.

Epilogue

I hope you enjoyed reading or perusing this book as much as I enjoyed writing it. I like to think that I have convinced you that basic understanding of semiconductor technology and devices is not out of the reach of any interested person and you do not need a PhD to understand how they work. Maybe I have convinced some younger (or older, why not?) students to go much deeper into the field or think seriously about a career in electronic engineering.

I have an undergraduate degree in philosophy with a minor in physics. The minor allowed me to go to Northwestern University to get a Masters in solid-state electronics and continue to my PhD at UCLA. I like to tell people that my two worst grades I got as I started my Masters program at Northwestern were in physical electronics and semiconductor devices. At the same time, I found the subject fascinating and I continued with it. I hope some of you may have similar experiences.

George Domingo

Appendix A
Useful Constants

A.1 Fundamental Physical Constants

Constants	Symbol	MKS	Units	CGS	Units
Bohr radius	a_o	5.30×10^{-11}	m	5.30×10^{-9}	cm
Boltzmann constant	k_b	1.38×10^{-23}	$J\,K^{-1}$	1.38×10^{-16}	$Erg\,K^{-1}$
Charge of electron	e	1.60×10^{-19}	C		
Electron mass	m_e	9.11×10^{-31}	kg	9.11×10^{-28}	g
Electron-volt (energy)	Ve	1.6×10^{-19}	J	1.6×10^{-12}	Erg
Permeability of free space	μ_o	1.26×10^{-6}	$H\,m^{-1}$	1.26×10^{-8}	$H\,cm^{-1}$
Permittivity of free space	ε_o	8.85×10^{-12}	$F\,m^{-1}$	8.85×10^{-14}	$F\,cm^{-1}$
Plank constant	h	6.63×10^{-34}	J s	6.63×10^{-27}	Ergs
Proton mass	m_c	1.67×10^{-27}	kg	1.67×10^{-24}	g
Rydberg constant	R_{oo}	1.1×10^{7}	m^{-1}		

Constants	Symbol	MKS	Units	CGS	Units
Speed of light	c	3.00×10^{8}	$\mathrm{m\,s^{-1}}$	3.00×10^{10}	$\mathrm{cm\,s^{-1}}$
Wavelength corresponding to 1 eV	λ	1.24×10^{-6}	m	1.24×10^{-4}	cm

A.2 Basic Units

Quantity	Name	Symbol
Electric current	Ampere	A
Length	Meter	m
Mass	Kilogram	kg
Temperature	Kelvin	K
	Centigrade	°C
Time	Second	s

A.3 Derived Units

Quantity	Name	Symbol	MKS
Acceleration		a	$\mathrm{m\,s^{-2}}$
Capacitance	Farad	F	$\mathrm{C\,V^{-1}}$
Conductance	Siemens	S	$\mathrm{A\,V^{-1}}$
Conductivity		σ	$\mathrm{A\,V^{-1}\,m^{-1}}$
Electric charge	Coulomb	C or Q	A s
Electric field		E	$\mathrm{V\,m^{-1}}$
Electrical potential	Volt	V or v	$\mathrm{kg\,m^{2}\,s^{-2}\,A^{-1}}$
Energy or work	Joule	J	$\mathrm{kg\,m^{2}\,s^{-2}}$
Force	Newton	N	$\mathrm{kg\,m\,s^{-2}}$

Quantity	Name	Symbol	MKS
Frequency	Hertz	Hz	s^{-1}
Inductance	Henry	H	$V\,s\,A^{-1}$
Luminous flux	Lumen	lm	$kg\,m^2\,s^{-3}$
Magnetic flux	Weber	Wv	$V\,s$
Power	Watt	W	$kg\,m^2\,s^{-3}$
Resistance	Ohms	Ω	$V\,A^{-1}$
Resistivity		ρ	$V\,A^{-1}\,m^{-1}$
Velocity		v	$m\,s^{-1}$
Wave number		v	m^{-1}
Wavelength		λ	m

Appendix B
Properties of Silicon

Property	Value	Units	Symbol
Atomic number	14		
Atomic weight	28.08	g	
Atoms cm^{-2}	5×10^{22}		
Cell volume	1.6×10^{-22}	cm^3	
Density	2.33	$g\,cm^{-3}$	
Density of atoms	5.0×10^{22}	cm^{-3}	
Dielectric constant (Si)	11.9		k(Si)
Dielectric constant (SiO_2)	3.9		$k(SiO_2)$
Diffusion constant for electrons	31	$cm^2\,s^{-1}$	D_n
Diffusion constant for holes	6.5	$cm^2\,s^{-1}$	D_p
Electron mobility (300 K)	1.4×10^3	$cm^2\,V\,s^{-1}$	μ_n
Energy gap	1.12	eV	E_g
Hole mobility (300°K)	471	$cm^2\,V\,s^{-1}$	
Index of refraction	3.42		
Intrinsic carrier concentration	1.45×10^{10}	cm^{-3}	n_i
Intrinsic resistivity (300°K)	6.36×10^4	Ω-cm	ρ_i
Lattice constant	5.43×10^{-8}	cm	a
Melting point	1415	°C	

Appendix C
List of Acronyms

A

A
Amplitude

A
Amperes

Å
Angstroms (10^{-10} m)

A
Area

a
Interatomic distance

AC
Alternating current

ALU
Arithmetic logic unit

AND
Logic circuit

As
Arsenic

B

B
Boron

Bit
Binary digit

BJT
Bipolar Junction transistor

Byte
Eight binary numbers

C

C
Capacitor (in Farads)

C
Constant

c
Speed of light

Cd
Cadmium

C_dS
Cadmium Sulfide

C_dT_e
Cadmium Telluride

cm
Centimeter

CPU
Central processing unit

D

d
Distance

DC
Direct current

DEMUX
Demultiplexer

D_n

Electron diffusion constant

D_p

Hole diffusion constant

DRAM

Dynamic random access memory

E

e

Electronic charge

E

Energy

E_A

Energy gap of an acceptor atom

E_D

Energy gap of a donor atom

EEPROM

Electrically erasable programable read-only memory

E_f

Fermi level

E_g

Energy gap between bands

EPROM

Erasable programable read-only memory

eV

Electron-volts

F

f

Frequency

F-D

Fermi-Dirac statistics

F(E)
Fermi function

FET
Field-effect transistor

FIR
Far infrared radiation

GaAs
Gallium Arsenide

GaP
Gallium Phosphate

Ge
Germanium

H

H
Henry, unit of inductance

h
Planck constant

HDL
Hardware description Language

Hg
Mercury

HgCaTe
Mercury Cadmium Tellurite

Hz
Hertz (unit of frequency)

I

I or I
Electrical current

I_B
Base current

I_C
Collector current

IC
Integrated circuit

I_{CEO}
Collector current at Base current = 0

I_E
Emitter current

I_n
Electron current

In
Indium

InAs
Indium Arsenide

I_{nD}
Electron diffusion current

I_{nE}
Electron drift current

InP
Indium Phosphate

InSb
Indium Antimonide

I_p
Hole current

i_{pD}
Hole diffusion current

i_{pE}
Hole drift current

ITO
Indium-tin-oxide

J

JFET
Junction field-effect transistor

K

k
Constant

K
Degrees Kelvin

k
Boltzmann constant

Kg
Kilograms

L

L
Inductance (in Henrys)

L
Length

l
Angular quantum number

LCD
Liquid crystal display

LED
Light emitting diode

M

m

Meter

m_e
Electron mass

MIPS
Millions of instructions per second

MIR
Mid infrared radiation

m_l
Magnetic quantum number

MOSFET
Metal oxide-semiconductor FET

MUX
Multiplexer

N

n
Levels of Bohr orbits

n
number of electrons

N
Number of turns in an inductor

N_A
Number of acceptor atoms

NAND
Inverted AND circuit

N_D
Number of donor atoms

n_i
Intrinsic number of electrons

n_i

Number of intrinsic electrons

NIR
Near infrared radiation

NOR
Logic circuit

n_w
Index of refraction

O

°C
Degrees centigrade

°F
Degrees Fahrenheit

OpAmp
Operational amplifier

Opcode
Operational code

OR
Logic circuit

P

P
Phosphorous

p
Number of holes

P
Power

p_i
Number of intrinsic holes

PROM
Programable read-only memory

Q

Q
: Electrical charge

qV
: Electron-Volt

R

R
: Resistance

R
: Rydberg constant

RAM
: Random access memories

r_b
: Transistor input resistance

R_C
: Collector resistance

R_E
: Emitter resistance

R_O
: Output resistance

ROM
: Read only memory

S

S
: Cross section of an inductor

s
: Seconds

SAM

Sequential access memories

Sb
Antimony

Si
Silicon

SOI
Silicon-on-insulator

SRAM
Static random access memory

T

T
Temperature

T_D
Timing delay

Te
Tellurium

TFET
Tunnel FET

TFT
Thin film transistor

U

USB
Universal serial buss

V

V
Volts

V or v
Electrical voltage

V
: Velocity

V_B
: Base voltage

V_C
: Collector voltage

V_{CB}
: Collector to base voltage

V_{CC}
: DC supply voltage

V_D
: Drain voltage

V_{DS}
: Drain to source voltage

V_E
: Emitter voltage

V_e
: External voltage

V_{EB}
: Emitter to base voltage

V_G
: Gate voltage

V_{GS}
: Gate to source voltage

V_i
: Internal voltage

V_{in}
: Input voltage

V_{out}
: Output voltage

W

W
Spectral radiation

W
Watts

W
Work function

W_M
Work function of metal

Word
4 bytes

W_S
Work function of semiconductor

X

X
Reluctance

X_C
Capacitance reluctance

XFIR
Extra far infrared radiation

X_L
Inductive reluctance

x_n
n-side transition region

XNOR
Logic circuit

x_p
p-side transition region

Z

Z
: Impedance

Greek letters

α
: Ratio of collector to emitter current

β
: Current gain

Δ
: Changing variable

ε_o
: Permittivity of free space

ε_r
: Relative Permittivity, Dielectric constant

θ
: Angle

ø
: Phase of sinusoidal wave

μm
: micrometers

μ
: Magnetic permittivity

λ
: Wavelength

μ_n
: Electron mobility

μ_o
: Susceptibility of permeability of free space

μ_p

Hole mobility

μ_r

Relative permeability

Ω

Ohms

ρ

Resistivity

ν

Wave number

ω

Frequency times 2π

Additional Reading and Sources

At the end of any technical book, authors add a list with a large number of books and articles related to the subjects they have covered in the book. This is not a textbook nor an academic research book, therefore I will try to be sparse and just mention those sources that either I used or that I think can be helpful in understanding the subject we reviewed.

By far the most useful tool for anyone who wants to learn a little more about any of the subjects I covered is *Wikipedia*. The quality of the information is quite good and up to date; sometimes a little too technical. The cross-referencing related subjects is also very useful.

Another source that I recommend is *YouTube*, which has some very informative and clear explanations that you can easily find by Googling the subject in which you are interested.

As far as books are concerned, a bibliography has to include books, obviously, so I list a few that amplify or clarify the topics in this book. There are thousands of books covering one portion or another of the topics I discussed. Take the recommendations here as books I have found useful; someone else would select other books.

I could not find any books in public libraries that, in simple terms, support, clarify or amplify the topics I cover in this book. You'll find books that tell you how to fabricate interesting electronic devices using transistors and OpAmps, but they do not explain how semiconductors work. One that does approach the topic in a very simple form is

Practical Electronics, Andy Cooper (John Murray Learning, 2016) This has 20 pages related to semiconductors but also many other simple explanations of electrical devices

College and university libraries have hundreds of books that cover the topics in this book. I mention here just a few. They are used as

an introduction to electrical engineering and assume you know or have an idea of calculus, for example:

Electrical Engineering, Principles and Applications, 4th edition, Allan R. Hambley (Pearson, Prentice Hall, 2008).

For semiconductor and device theory, I recommend some old, classic books that are still available that are quite simple and clear. Additionally, they show how little the fundamental theory of semiconductors has changed. All require calculus, but many can be understood by skipping the equations:

An Introduction to Junction Transistor Theory, Robert D. Middlebrook (John Wiley & Sons, 1957) This has one qualitative chapter with simple math followed by a quantitative one, still available after so many years.

Physics and Technology of Semiconductor Devices, Andrew S. Grove (John Wiley & Sons, 1967) Andrew Grove not only was a great scientist but he became the CEO of Intel. An excellent book.

Physics of Semiconductor Devices (many editions), Simon M. Sze (John Wiley & Sons, 1969) A real classic and complete coverage of the theory of semiconductor devices.

Semiconductor Devices, Basic Principles, Jasprit Singh (Wiley, 2001) Very rigorous and complete. Requires calculus, but not very mathematical.

Integrated Microelectronic Devices, Jesus A. del Alamo (Pearson, 2018) Covers transistor uses and performance.

Other books put more emphasis on modern electronics and integrated circuits and less on the physics behind the devices:

Design of Operational Amplifiers and Analog Integrated Circuits, Sergio Franco (MacGraw-Hill, 2015) Very good and more recent.

CMOS Digital Integrated Circuits, Analysis and Design, Sung-Mo Kang, Yusuf Leblebici, and Chulwo Kim (McGraw-Hill, 2015) Concentrates on CMOS devices, fabrication, and uses in logic circuits

CMOS Digital Integrated Circuits: A first cover, C. Hawkins, J. Segura, and P. Zarkesh-Ha (Scitech, 2013) Very good and complete coverage, including devices based on Boolean algebra using lots of examples and simple math.

There are many books that talk about computers. Even though they may be very technical, they use very little math. Here are a few:

The Magda Guide to Microprocessors, Michio Shibuya, Tarashi Tonagi, and Office Sawa (No Starch Press, 2017) A fun cartoon book explaining computers that is easy to follow and very complete and thorough

Design of Digital Computers, an Introduction, Hans W. Gschwind and Edward J. NcCluskey (Springer-Verlag, 1975) Quite basic, little math, old but still good information.

Digital Design and Computer Architecture, David Harris and Sarah Harris (Morgan Kaufmann, 2007) Two volumes

Introduction to Embedded Systems: A Cyber Physical Systems Approach, Edward Lee and Sanjit Seshia (LeeSeshia.org, 2011) Very good with lots of examples, emphasizes the programing of these systems

What Every Engineer Should Know About Developing Real-Time Embedded Products, 2nd edition, Kim R. Fowler (CRC Press, 2017) Using specific case studies this book explains the uses of embedded systems very well.

The Microprocessor: A Biography, Michael S. Malone (Springer-Verlog/EROS, 1995) Even though it is old, this has a nice simple historical coverage of the microprocessors

Inside the Machine: An Illustrated Introduction to Microprocessors and Computer Architecture, Jon Stokes (No Starch Press, 2007) Very complete and practically no math

Fundamentals of Modern Electronic Devices, 2nd edition, Yuan Taur and Tak Ning (Cambridge University Press, 2014) Very good explanations on how memories work.

Almost all books on semiconductors have some chapters on how integrated circuit devices are fabricated. Books that emphasize the fabrication aspect of semiconductors include the following:

Fabrication Engineering at the Micro and Nanoscale, Stephen A. Campbell (Oxford University Press, 2008) Very detailed explanations with lots of references.

Integrated Circuit Packaging, Assembly and Interconnects, William J. Greing (Springer, 2006) The first couple of chapters discusses processing and the rest of the book is on packaging.

Microlithography: Process technology for IC fabrication, David Elliott (McGraw Hill, 1986). Old but very complete and readable book. No math and covers all the processing steps.

Fundamentals of Microfabrication, 2nd edition, Mark Madau (CRC Press, 2002) Has good explanations and figures.

Books on new technologies are quite technical and mathematical, but again some of the explanations are clear. Here are some:

The Quantum Story: A history in 40 moments, Jim Baggott (Oxford University Press, 2011) Very easy to read, entertaining, and informative.

Principles of Quantum Computation and Information, Guiliano Beneti, Giulio Casati, and Giuliano Strini (World Scientific, 2004) In two volumes, very complete, quite mathematical, but of course is quantum mechanical!

Bio-inspired and Nanoscale Integrated Computing, edited by Mary Mehrnoosh Eshaghian-Wilner (Wiley, 2009) The first four sections are quite instructive, but the rest are on rather specialized topics

Introduction to Nanoelectronics: Science, Nanotechnology, and Applications, Vladimir Mitin, Viatcheslav Kochelap, and Michael Stroscio (Cambridge University Press, 2008) Very mathematical but explanations are clear

For optoelectronics I suggest

Photonics Essentials, Thomas P. Pearsall (McGraw-Hill, 2003) Covers most of the topics with relatively simple math.

Optics of Liquid Crystal Displays, Pochi Yeh and Claire Gu (John Wiley & Sons, 2010) First chapter has lots of detail that they expand in the subsequent chapters

Understanding Lasers, Jeff Hecht (John Wiley & Sons, 2008)

More Things in the Heavens, Michael Werner and Peter Eisenhardt (Princeton University Press, 2019) Covers the latest uses of infrared detectors in astronomy.

Finally, some books on the future of semiconductors:

Future Trends in Microelectronics: Up to Nano Creek, edited by Serge Lurvi, Jimmy Xu, and Alen Zaslavsky (Wiley, 2007) Although more than 10 years old, this has good coverage of where the technology is going

Index

There are many concepts, ideas, subjects and words that appear repeatedly in the text. The words "semiconductors", "Conduction band" and "pn-junction", for example, appear multiple times in the text. I list in the index the first time that the concept appears and use bold numbers to indicate pages where there is a relevant explanation of the concept

a

b

c

d

e

f

g

h

i

j

k

l

m

n

o

p

q

r

S

t

u

v

w

x

y

z

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.