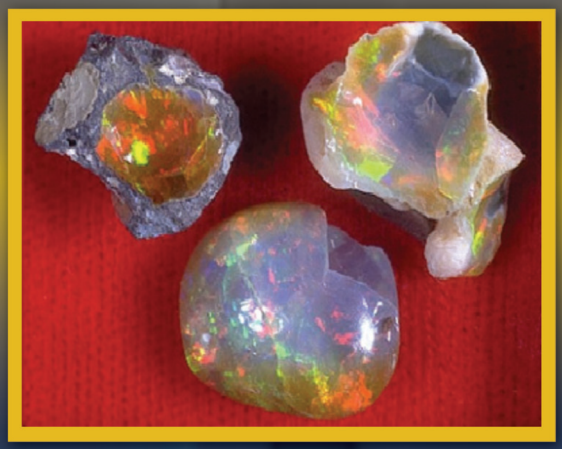
Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology



Marc J. Madou

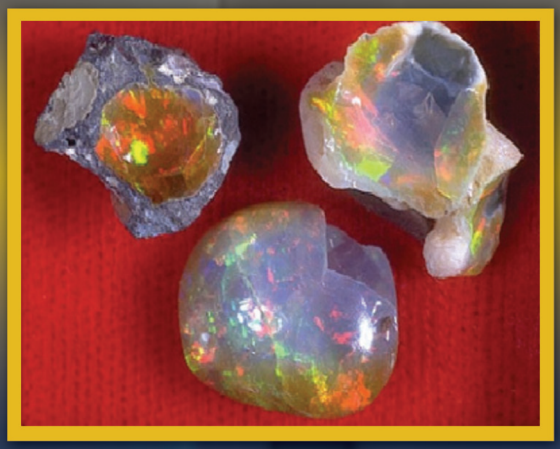
CRC Press
Taylor & Francis Group

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME I

Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology



Marc J. Madou

CRC Press
Taylor & Francis Group

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME I

Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Third Edition

Fundamentals of MICROFABRICATION AND NANOTECHNOLOGY

VOLUME I

Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology

Marc J. Madou

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper

International Standard Book Number: 978-1-4200-5511-5 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

I dedicate this third edition of Fundamentals of Microfabrication to my family in the US and in Belgium and to all MEMS and NEMS colleagues in labs in the US, Canada, India, Korea, Mexico, Malaysia, Switzerland, Sweden and Denmark that I have the pleasure to work with. The opportunity to carry out international research in MEMS and NEMS and writing a textbook about it has been rewarding in terms of research productivity but perhaps even more in cultural enrichment. Scientists have always been at the frontier of globalization because science is the biggest gift one country can give to another and perhaps the best road to a more peaceful world.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Roadmap

In *Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology* we lay the foundations for a qualitative and quantitative theoretical understanding of micro- and nanoelectromechanical systems, i.e., MEMS and NEMS. In integrated circuits (ICs), MEMS, and NEMS, silicon (Si) is still the substrate and construction material of choice. A historical note about the ascent of silicon, MEMS and NEMS is the topic of Chapter 1. The necessary solid-state physics background to understanding the electronic, mechanical, and optical properties of solids relied on in ICs, MEMS and NEMS is covered in Chapters 2–5. Many important semiconductor devices are based on crystalline materials because of their reproducible and predictable electrical properties. We cover crystallography in Chapter 2. The ultimate theory in modern physics today to predict physical, mechanical, chemical, and electrical properties of atoms, molecules, and solids is quantum mechanics. Quantum mechanics and the band theory of solids are presented in Chapter 3. The relevance of quantum mechanics in the context of ICs and NEMS cannot be underestimated, and the profound implications of quantum physics for nanoelectronics and NEMS are a recurring topic throughout this book. Given the importance of single-crystal Si (SCS) for IC, MEMS, and NEMS applications, we analyze silicon crystallography and band structure in more detail in Chapter 4. This chapter also elucidates all the single-crystal Si properties that conspired to make Si so important in electronic, optical, and mechanical devices that one might rightly call the second half of the 20th century the Silicon Age. Photonics, treated in Chapter 5, involves the use of radiant energy and uses photons the same way that electronic applications use electrons. We review the distinctive optical properties of bulk 3D metals, insulators, and semiconductors and summarize effects of electron and photon confinement in lower-dimensional structures. We show how evanescent fields on metal surfaces enable the guiding of light below the diffraction limit in plasmonics. Plasmonics is of growing importance for use in submicron lithography, near-field optical microscopy, enhancement of light/matter interaction in sensors, high-density data storage, and highly integrated optic chips. We also delve into the fascinating new topic of metamaterials, man-made structures with a negative refractive index, and explain how this could make for perfect lenses and could change the photonic field forever. In Chapter 6 we introduce fluidics, compare various fluidic propulsion mechanisms, and discuss the influence of miniaturization on fluid behavior. Given the high level of interest, fluidics for miniaturized analytical equipment is covered in this chapter as well. Chapter 7 combines a treatise on electrochemical and optical analytical processes whose implementation is often attempted in miniaturized components and systems.

Note to the Reader: *Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology* was originally composed as part of a larger book that has since been broken up into three separate volumes. *Solid-State Physics, Fluidics, and Analytical Techniques in Micro- and Nanotechnology* represents the first volume in this set. The other two volumes include *Manufacturing Techniques for Microfabrication and Nanotechnology* and *From MEMS to Bio-NEMS: Manufacturing Techniques and Applications*. Cross-references to these books appear throughout the text and will be referred to as Volume II and Volume III, respectively. The interested reader is encouraged to consult these volumes as necessary.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Author

Dr. Madou is the Chancellor's Professor in Mechanical and Aerospace Engineering (MEA) at the University of California, Irvine. He is also associated with UC Irvine's Department of Biomedical Engineering and the Department of Chemical Engineering and Materials Science. He is a Distinguished Honorary Professor at the Indian Institute of Technology, Kanpur, India and a World Class University Scholar (WCU) at UNIST in South Korea.

Dr. Madou was Vice President of Advanced Technology at Nanogen in San Diego, California. He specializes in the application of miniaturization technology to chemical and biological problems (bio-MEMS). He is the author of several books in this burgeoning field he helped pioneer both in academia and in industry. He founded several micromachining companies.

Many of his students became well known in their own right in academia and through successful MEMS start-ups. Dr. Madou was the founder of the SRI International's Microsensor Department, founder and president of Teknekron Sensor Development Corporation (TSDC), Visiting Miller Professor at UC Berkeley, and Endowed Chair at the Ohio State University (Professor in Chemistry and Materials Science and Engineering).

Some of Dr. Madou's recent research work involves artificial muscle for responsive drug delivery, carbon-MEMS (C-MEMS), a CD-based fluidic platform, solid-state pH electrodes, and integrating fluidics with DNA arrays, as well as label-free assays for the molecular diagnostics platform of the future.

To find out more about those recent research projects, visit http://www.biomems.net.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Acknowledgments

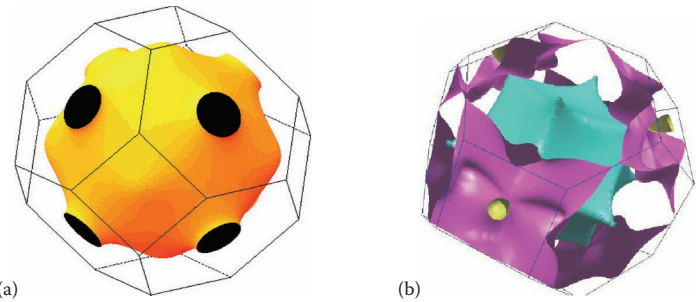
I thank all of the readers of the first and second editions of Fundamentals of Microfabrication as they made it worthwhile for me to finish this completely revised and very much expanded third edition. As in previous editions I had plenty of eager reviewers in my students and colleagues from all around the world. Students were especially helpful with the question and answer books that come with the three volumes that make up this third edition. I have acknowledged reviewers at the end of each chapter and students that worked on questions and answers are listed in the questions sections. The idea of treating MEMS and NEMS processes as some of a myriad of advanced manufacturing approaches came about while working on a WTEC report on International Assessment Of Research And Development In Micromanufacturing (http://www.wtec.org/micromfg/report/Micro-report.pdf). For that report we travelled around the US and abroad to visit the leading manufacturers of advanced technology products and quickly learned that innovation and advanced manufacturing are very much interlinked because new product demands stimulate the invention of new materials and processes. The loss of manufacturing in a country goes well beyond the loss of only one class of products. If a technical community is dissociated from manufacturing experience , such as making larger flat-panel displays or the latest mobile phones, such communities cannot invent and eventually can no longer teach engineering effectively. An equally sobering realization is that a country might still invent new technologies paid for by government grants, say in nanofabrication, but not be able to manufacture the products that incorporate them. It is naïve to believe that one can still design new products when disconnected from advanced manufacturing: for a good design one needs to know the latest manufacturing processes and newest materials. It is my sincerest hope that this third edition motivates some of the brightest students to start designing and making things again rather than joining financial institutions that produce nothing for society at large but rather break things.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Introduction

MEMS and NEMS Foundations

Miniaturization science is the science of making very small things. In top-down micro- and nanomachining, one builds down from large chunks of material; in bottom-up nanochemistry, one builds up from smaller building blocks. Both require a profound understanding of the intended application, different manufacturing options, materials properties, and scaling laws. The resulting three-dimensional structures, ranging in size from subcentimeters to subnanometers, include electronics, photonics, sensors, actuators, micro- and nanocomponents, and micro- and nanosystems.



(a) (b)

(a) Copper Fermi surface—FCC sixth band with eight short necks touching the eight hexagonal zone faces. (Fermi surface database at http://www.phys.ufl.edu/fermisurface.) (b) Platinum Fermi surface—FCC fourth, fifth, and sixth bands. (Fermi surface database at http://www.phys.ufl.edu/fermisurface.)

No one behind, no one ahead. The path the ancients cleared has closed. And the other path, everyone's path, easy and wide, goes nowhere. I am alone and find my way.

Dharmakirti
(7th century India)

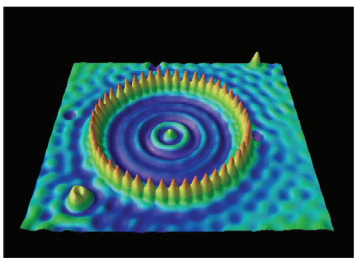
Introduction

Introduction

In Volume I, we lay the foundations for a qualitative and quantitative understanding of micro- and nanoelectromechanical systems, i.e., MEMS and NEMS. In integrated circuits (ICs), MEMS, and NEMS, silicon (Si) is still the substrate and building material of choice. A historical note about the history of the ascent of silicon, MEMS, and NEMS is the topic of Chapter 1.

The necessary solid-state physics background of electronic, mechanical, and optical properties of solids relied on in MEMS and NEMS is covered in Chapters 2–5. Solid-state physics is the study of solids. A major part of solid-state physics is focused on crystals because the periodicity of atoms in a crystal facilitates mathematical modeling, but more importantly because crystalline materials often have electrical, optical, or mechanical properties that can be easier exploited for engineering purposes. In Chapter 2, we detail crystalline materials in which atoms are arranged in a pattern that repeats periodically in three dimensions. The materials covered here prepare the reader for Chapter 3, which explains the band theory of solids based on quantum mechanics. The relevance of quantum mechanics in the context of ICs and NEMS cannot be underestimated, and the profound implications of quantum physics for nanoelectronics and NEMS is a recurring topic throughout this book. This is followed in Chapter 4 by a description of the single-crystal Si band structure, the growth of single crystals of Si, its doping, and oxidation. In this chapter, we also review the single-crystal Si properties that conspired to make Si so important in electronic, optical, and mechanical devices so that one might rightly call the second half of the 20th century the Silicon Age. Although the emphasis in this book is on nonelectronic applications of miniaturized devices, we briefly introduce different types of diodes and two types of transistors (bipolar and MOSFET). In Chapter 5, we introduce photonics. We compare electron and photon propagation in materials and contrast electron and photonic confinement structures and the associated evanescent wave phenomena. We also delve into the fascinating new topic of metamaterials, artificially engineered materials possessing properties (e.g., optical, electrical) that are not encountered in naturally occurring ones. An introduction to diode lasers, quantum well lasers, and quantum cascade lasers concludes the photonics section.

Fluidics and electrochemical and optical analytical techniques are important current applications of MEMS and NEMS. In Chapter 6 we introduce fluidics, compare various fluidic propulsion mechanisms, and discuss the influence of miniaturization on fluid behavior. Given the current academic and industrial interest, fluidics in miniaturized analytical equipment is detailed separately at the end of this chapter. Chapter 7 combines a treatise on



STM image showing standing waves in a 2D electron gas trapped in a "quantum corral" made by positioning Fe atoms on a Cu (111) surface.[1]

electrochemical and optical analytical techniques. Using sensor examples, we introduce some of the most important concepts in electrochemistry, i.e., the electrical double layer, potentiometry, voltammetry, two- and three-electrode systems, Marcus' theory of electron transfer, reaction rate- and diffusion rate-controlled electrochemical reactions, and ultramicroelectrodes. Many researchers use MEMS and NEMS to miniaturize optical components or whole instruments for absorption, luminescence, or phosphorescence spectroscopy. Optical spectroscopy is concerned with the production, measurement, and interpretation of electromagnetic spectra arising from either emission or absorption of radiant energy by matter. The sensitivity of these optical sensing techniques and the analysis of how amenable they are to miniaturization (scaling laws) are also compared herein. Chapter 7 ends with a comparison of the merits and problems associated with electrochemical and optical measuring techniques.

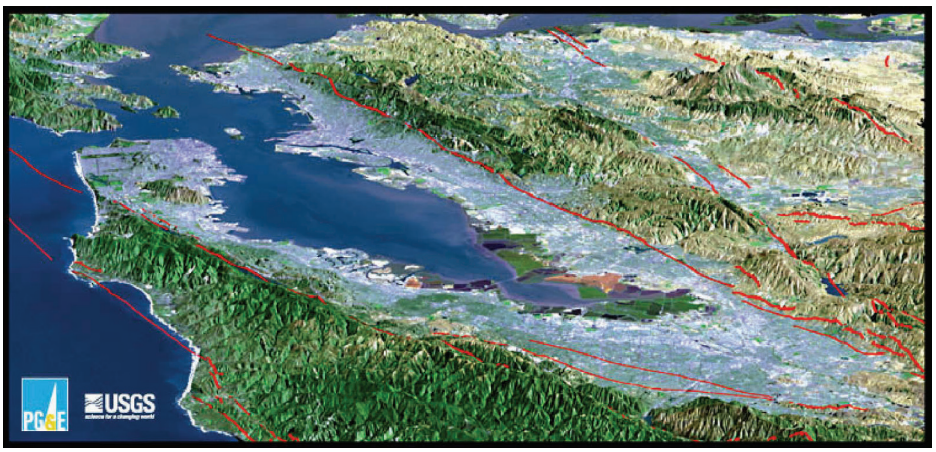
Reference

1. Crommie, M. F., C. P. Lutz, and D. M. Eigler. 1993. Confinement of electrons to quantum corrals on a metal surface. *Science* 262:218–20.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Historical Note: The Ascent of Silicon, MEMS, and NEMS



Silicon Valley is the only place on Earth not trying to figure out how to become Silicon Valley (Robert Metcalfe, father of the ethernet). (From PG&E and USGS.)

Outline

Silicon in Integrated Circuits

In 1879, William Crookes recounted his experiments on passing electric discharges through an evacuated glass tube for the Royal Society, thus describing the first cathode ray tube (CRT). Four years later, Thomas Alva Edison and Francis Upton discovered the "Edison effect." They introduced a metal plate into an incandescent electric light bulb (invented by Edison in 1879) in an attempt to keep the bulb from turning black (Figure 1.1). It did not work, but they discovered that there was a current between the lighted filament and the separate metal plate when the plate was positively charged but not when it was negatively charged. This led Edison and Upton to stumble on the basic principle of the operation of the vacuum tube (rectification!).

The first diode tube we owe to John Fleming, who, in 1904, filed a patent for a "valve" vacuum tube, also called a *Fleming valve* or *Fleming diode*.

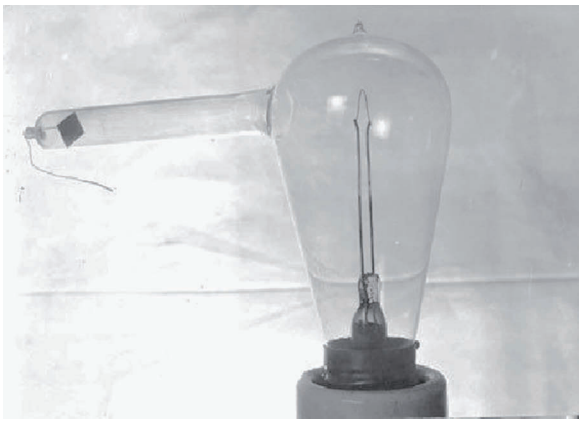


FIGURE 1.1 Edison bulb used to demonstrate the "Edison effect."

Early researchers thought of electricity as a kind of fluid, leading to the inherited jargon such as current, flow, and valve. Fleming recognized the importance of Edison and Upton's discovery and demonstrated it could be used for the rectification of alternating currents. Interestingly, Fleming, at first, tried to get reliable rectification from single-crystal rectifiers used in crystal set radios (Figure 1.2), but could never get them to work well enough, so he switched to tubes!

J.J. Thomson, in 1887, convincingly showed that an electrical current was really an electron flow, and Fleming could explain the rectification in his diode tube as electrons boiling of the heated filament and flowing to the metal plate (thermionic emission). Because the plate was not hot enough to emit electrons, no current could go in the opposite direction. Thus, the Edison effect always produced direct

FIGURE 1.2 Early crystal set radio with a galena (lead sulphide) and the "cat's whisker" (the small coil of wire) that was used to make contact with the crystal.

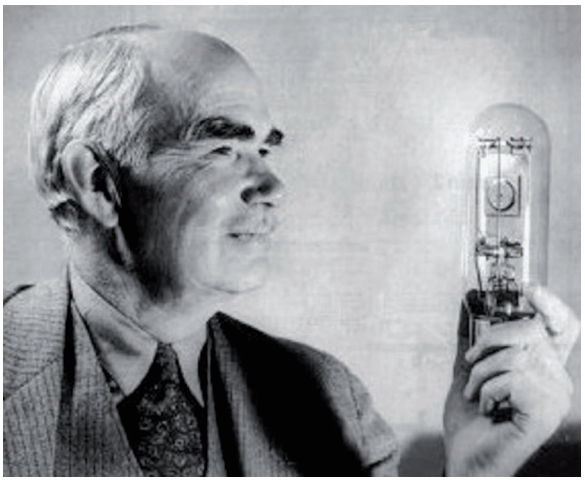


FIGURE 1.3 Lee De Forest in 1906 with the Audion, the first triode.

current only. In 1906, an American scientist, Lee De Forest (Figure 1.3), invented the vacuum tube amplifier or triode based on the two-element vacuum tube invented by Fleming. De Forest, reportedly a tireless self-promoter, added an electrode—the *grid*—to the Fleming diode, and inserted it between the anode and the cathode. With this grid the diode became an active device, i.e., it could be used for the amplification of signals (say, for example, in radios) and as a switch (for computers). Hence, the amplifying vacuum tube, the ancestor of the transistor, was born. A gate in a dam controls huge amounts of flowing water with relatively small movements. Similarly, a small signal applied to the grid controls the much larger signal between anode and cathode.

Vacuum tubes—miniature particle accelerators—dominated the radio and television industries until the 1960s, and were the genesis of today's huge electronics industry. However, tubes were fragile, large, very power hungry, and costly to manufacture. The industry needed something better. That today's world is largely electronic—e.g., automobiles, home appliances, even books, writing tablets, and tally sheets—is because of solid-state electronics,* not vacuum tubes.

It is true, albeit unfortunate, that World War II and the subsequent Cold War era is what spurred research and development in solid-state electronic devices. Human foibles led to faster development of

* Based on or consisting chiefly or exclusively of semiconducting materials, components, and related devices.

RAdio **D**etecting **A**nd **R**anging or RADAR, **SO**und **N**avigation **A**nd **R**anging or SONAR, and many other technological innovations. As we all know, the list of innovations made to feed human aggression did not abate. Alan Turing led the team in England that in 1943 built the Colossus coding and deciphering machine. The Colossus was a special-purpose computer used to break the German code ULTRA, encrypted using ENIGMA machines. Breaking the German code was one of the keys to the success of the D-Day invasion. The Harvard Mark I and later II, III, and IV were general-purpose electromechanical calculators (sponsored by the U.S. Navy) to compute artillery and navigation tables—the same purpose intended, 100 years earlier, by Babbage for his analytical engine (Figure 1.4).

John Mauchly and Presper Eckert started work on the first electronic computer, the ENIAC (Electronic Numerical Integrator and Computer), at the University of Pennsylvania in 1943. The ENIAC, having been a secret during the war, was unveiled in Philadelphia in 1946. This computer featured 17,468 vacuum tubes used as switches and consumed 174 kW of power, enough to light 10 homes! Several tubes would fail every day until the engineers decided to never turn off the machine. This increased the average time until a tube would fail to 2 days. ENIAC was designed to calculate munition trajectory tables for the U.S. Army. It was U-shaped, 25 m long, 2.5 m high, 1 m wide, and weighed more than 30 tons (see Figure 1.5a). Programming was done by plugging cables and setting switches. By the mid-1970s, identical ENIAC functions could be achieved by a 1.5- × 1.5-cm silicon die, and the original Pentium processor, if fabricated using ENIAC technology, would cover more than 10 square miles.

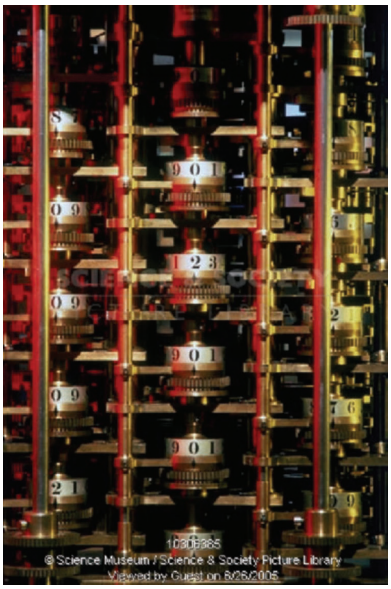


FIGURE 1.4 Charles Babbage (1791–1871) first conceived the idea of an advanced calculating machine to calculate and print mathematical tables in 1812. It was a decimal digital machine, with the value of a number being represented by the positions of toothed wheels marked with decimal numbers.

One of ENIAC's heirs was a computer called the UNIVAC (Universal Automatic Computer), considered by most historians to be the first commercially successful digital computer (Figure 1.5b). First constructed by the Eckert-Mauchly Computer Corporation (EMCC), it was taken over by Sperry-Rand. At 14.5 ft. long, 7.5 ft. high, and 9 ft. wide, the UNIVAC, priced at $1 million, was physically

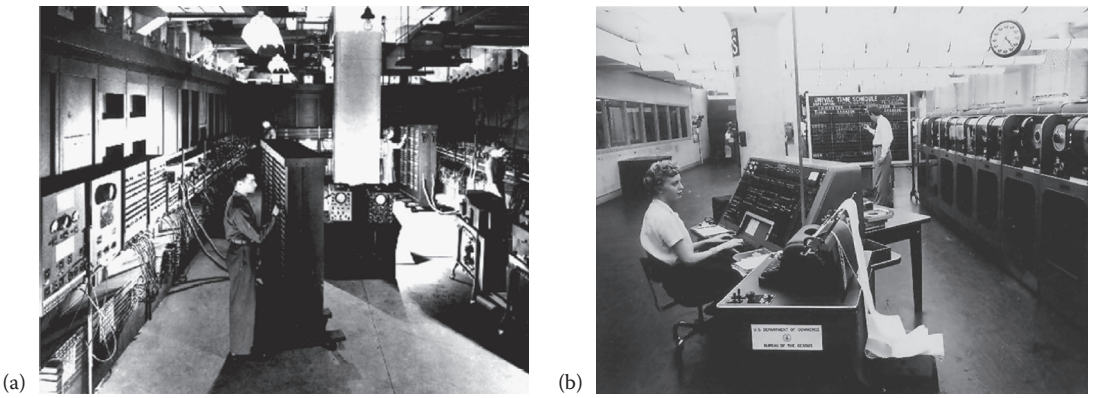


FIGURE 1.5 Electronic Numerical Integrator and Computer (ENIAC), the world's first large-scale, general-purpose electronic computer (a), and the UNIVAC, the first commercial computer (b). First-generation computers based on vacuum tubes.

FIGURE 1.6 Von Neumann in his living room. (Photograph by Alan Richards hanging in Fuld Hall, Institute for Advanced Study, Princeton, NJ. Courtesy of the Archives of the Institute for Advanced Study.)

FIGURE 1.7 The first point-contact germanium bipolar transistor. Notice the paper clip! Roughly 50 years later, electronics accounted for 10% ($4 trillion) of the world's aggregate GDP.

smaller than ENIAC but more powerful. ENIAC and UNIVAC constitute first-generation computers based on vacuum tubes.

It was the concept of the stored program, invented by John von Neumann in 1945 (Figure 1.6); the magnetic core memory, invented by An Wang at Harvard and used in grids by J.W. Forrester and colleagues at MIT for random access memory (RAM); and William Shockley's transistor, based on transistors for switches instead of tubes, that would make a second generation of computers possible, thus starting the computer revolution.

The year 1940 gave rise to an important milestone in solid-state electronics history with the invention of a silicon-based solid-state p-n junction diode by Russell Ohl at Bell Labs.[1] This device, when exposed to light, produced a 0.5 V across the junction and represented the first Si-based solar cell. Bell Labs, in 1945, established a group charged with developing an alternative to the vacuum tube. Led by Shockley (1910–1989) and including John Bardeen (1908–1991) and Walter Brattain (1902–1987), in 1947 the group made an odd-looking device consisting of semiconducting germanium (Ge), gold strips, insulators, and wires, which they called a transistor (subject of U.S. Patent #2,524,035 [1950][2]) (Figure 1.7; notice the paper clip!). For this invention Shockley, Bardeen, and Brattain were awarded the 1956 Nobel Prize for Physics (Bardeen went on to claim a second Nobel Prize for Physics in 1972 for superconductivity). Bardeen called Ohl's junction diode fundamental to the invention of the transistor. Brattain was the unassuming experimentalist, Bardeen the theorist, and Shockley* the inventor and leader (Figure 1.8).

This trio thus succeeded in creating an amplifying circuit using a point-contact bipolar transistor that *trans-ferred* resistance (hence *transistor*). Two

FIGURE 1.8 Shockley (seated), Bardeen, and Brattain.

* Unfortunately, Shockley became associated with racist ideas and briefly pursued a U.S. Senate seat (as a Republican).

wires made contact with the germanium crystal near the junction between the p- and n-zones just like the "cat whiskers"* in a crystal radio set. A few months later, Shockley devised the junction transistor, a true solid-state device that did not need the whiskers of the point-contact transistor (see also Figure 1.2). Junction transistors were much easier to manufacture than point-contact transistors, and by the mid-1950s the former had replaced the latter in telephone systems. G. Teal and J.B. Little, also from Bell Labs, were able to grow large single crystals of germanium by 1951, which led to the start of commercial production of germanium transistors in the same year. Christmas 1954 saw the first transistor radio (the Regency TR-1) built by Industrial Development Engineering Associates, which sold for $49.95 (Figure 1.9). This radio featured four germanium transistors from Texas Instruments. Although germanium was used in early transistors, by the late 1960s silicon, because of its many advantages, had taken over.

Silicon has a wider bandgap (1.1 eV for Si vs. 0.66 eV for Ge), allowing for higher operating temperatures (125–175°C vs. 85°C), a higher intrinsic resistivity (2.3×10^5 Ω cm vs. 47 Ω cm), and a better native oxide (SiO_2 vs. GeO_2 [water soluble!]).

FIGURE 1.9 Movie producer mogul Michael Todd (husband of Elizabeth Taylor in the mid-1950s) placed Regency TR-1s in gift books to commemorate his movie *Around the World in 80 Days*. The one pictured was for Shirley MacLaine.

* A cat whisker is a piece (often springy) of pointed metal wire.

The latter results in a higher-quality insulator that protects and "passivates" underlying circuitry, helps in patterning, and is useful as a mask for dopants. Finally, silicon is cheaper and much more abundant (sand!) than germanium. Second-generation computers relied on transistors instead of vacuum tubes for switches (logic gates). In recent years, germanium is making a comeback based mostly on its higher carrier mobility (three times higher than silicon-based ones), of great interest for faster circuitry, and because Ge has a lattice constant similar to GaAs, making it easier to integrate GaAs optical components with Ge-CMOS circuits.

Transistors perform functions similar to vacuum tubes, but they are much smaller, cheaper, less power hungry, and more reliable. Michael Riordan and Lillian Hoddeson's *Crystal Fire* gives, in the author's opinion, one of the best popular accounts of the invention of the transistor.[3]

The honeymoon with the transistor was quickly over; by the second half of the 1950s, new circuits on the drawing board were so big and complex that it was virtually impossible to wire that many different parts together reliably. A circuit with 100,000 components could easily require 1 million, mostly manual, soldering operations that were time consuming, expensive, and inherently unreliable. The answer was the "monolithic" idea, in which a single bloc of semiconductor is used for all the components and interconnects, invented by two engineers working at competing companies: Jack Kilby at Texas Instruments (Figure 1.10) and Robert Noyce (Figure 1.11) at Fairchild Semiconductor.

In 1958, Jack Kilby at Texas Instruments formed a complete circuit on a piece of germanium, landing U.S. Patent #3,138,743 (1959). His circuit was a simple IC oscillator with three types of components: transistors, resistors, and capacitors (Figure 1.12). Kilby got his well-deserved Nobel Prize for this work only in 2000. Technological progress and engineering feats are not often awarded a Nobel Prize, and if awarded at all they are often belated or controversial (see Kary Mullis, Nobel Laureate Chemistry 1993 for the invention of PCR).

Robert Noyce—Mr. Intel (Integrated Electronics)—then at Fairchild, introduced, with Jean Horni, planar technology, wiring individual devices together

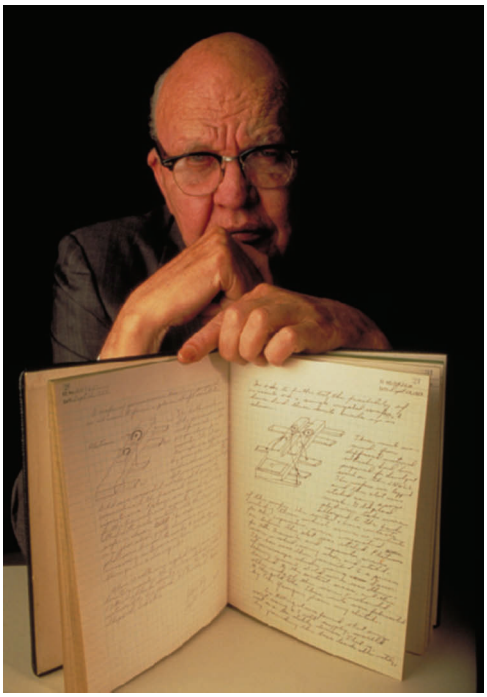


FIGURE 1.10 Jack Kilby with notebook. (TI downloadable pictures.)

on a silicon wafer surface. Noyce's "planar" manufacturing, in which all the transistors, capacitors, and resistors are formed together on a silicon chip with the metal wiring embedded on the silicon, is still used today. By 1961, Fairchild and Texas Instruments had devised methods whereby large

FIGURE 1.11 Robert Noyce in 1990.

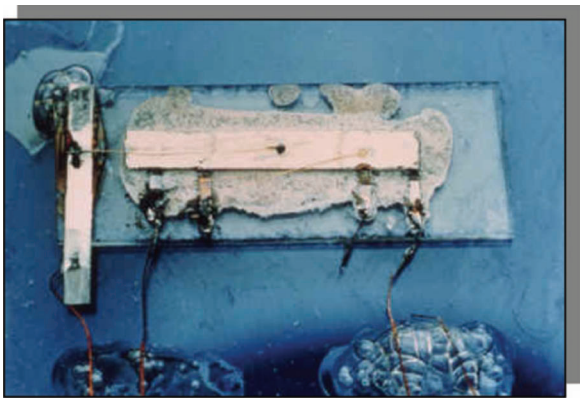


FIGURE 1.12 The first integrated circuit (germanium) in 1958 by Jack S. Kilby at Texas Instruments contained five components of three types: transistors, resistors, and capacitors.

numbers of transistors were produced on a thin slice of Si—and IC production on an industrial scale took off. The transistors on ICs were not the bipolar type but rather field effect transistor devices. The concept of a field effect transistor (FET) was first proposed and patented in the 1930s; however, it was the bipolar transistor that made it first to commercial products. Shockley resurrected the idea of the FET in the early 1950s, but it took until 1962 before a working FET was fabricated. These new FETs proved to be more compatible with both IC and Si technology.

Integrated circuits made not only third-generation computers possible but also cameras, clocks, PDAs, RF-IDs, etc. The National Academy of Sciences declared ICs the progenitor of the "Second Industrial Revolution," and Jerry Sanders, founder of Advanced Microdevices, Inc., called ICs the crude oil of the 1980s. A very well-written popular account of the invention of the IC is T.R. Reid's *The Chip: How Two Americans Invented the Microchip and Launched a Revolution*.[4]

Robert Noyce, Gordon Moore, and Andrew Grove left Fairchild to start Intel in 1968 with the aim of developing random access memory (RAM) chips. The question these inventors wanted answered was this: since transistors, capacitors, and resistors can be put on a chip, would it be possible to put a computer's central processor unit (CPU) on one? The answer came swiftly; by 1969 Ted Hoff had designed the Intel 4004, the first general-purpose 4-bit microprocessor. The Intel 4004 microprocessor was a 3-chip set with a 2-kbit read-only memory (ROM) IC, a 320-bit RAM

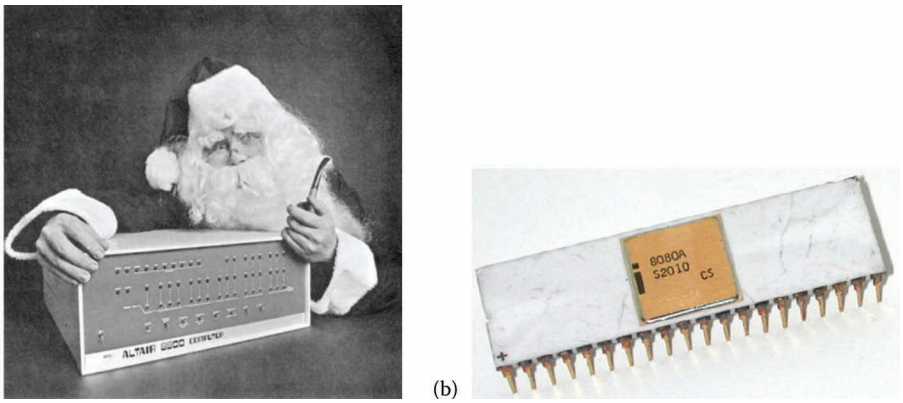


FIGURE 1.13 (a) The Altair 8800 computer and (b) the Intel 8080 microprocessor.

IC, and a 4-bit processor, each housed in a 16-pin dual in-line package (DIP). The processor, made in a 10-µm silicon gate pMOS process, contained 2,250 transistors and could execute 60,000 operations per second on a die size of 13.5 mm^2. It came on the market in 1971, giving rise to the fourth generation of computers based on microprocessors and the first personal computer (PC). The era of a computer in every home—a favorite topic among science fiction writers—had arrived!

The first desktop-size PC appeared in 1975, offered by Micro Instrumentation Telemetry Systems (MITS) as a mail-order computer kit. The computer, the Altair 8800, named after a planet on a *Star Trek* episode, retailed for $397. It had an Intel 8080 microprocessor, 256 bytes of memory (not 256K), no keyboard, no display, and no auxiliary storage device, but its success was phenomenal, and the demand for the microcomputer kit was overwhelming (Figure 1.13). Scores of small entrepreneurial companies responded to this demand by producing computers for the new market. In 1976, Bill Gates, Paul Allen, and Monte Davidoff wrote their first software program for the Altair—a BASIC (Beginners All-purpose Symbolic Instruction Code) interpreter (a high-level language translator that converts individual high-level computer language program instructions [source code] into machine instructions).

The first major electronics firm to manufacture and sell personal computers, Tandy Corporation (Radio Shack), introduced its computer model (TRS-80) in 1977. It quickly dominated the field because of the combination of two attractive features: a keyboard and a cathode ray display terminal. It was also popular because it could be programmed, and the user was able to store information by means of a cassette tape. In 1976, Steve Wozniak, who could not afford an Altair, built his own computer using a cheaper microprocessor and adding several memory chips. As a circuit board alone, it could do more than the Altair. Wozniak and Steve Jobs called it Apple I, and Jobs took on the task of marketing it while Wozniak continued with improvements (Figure 1.14). By 1977, Wozniak had built the Apple II and quit his day job. The Apple II had 16–64K RAM and secondary memory storage in the shape of a cassette tape or a 5.25-in. floppy disk drive and cost $1,300. At that time, Wozniak and Jobs formed Apple Computer,

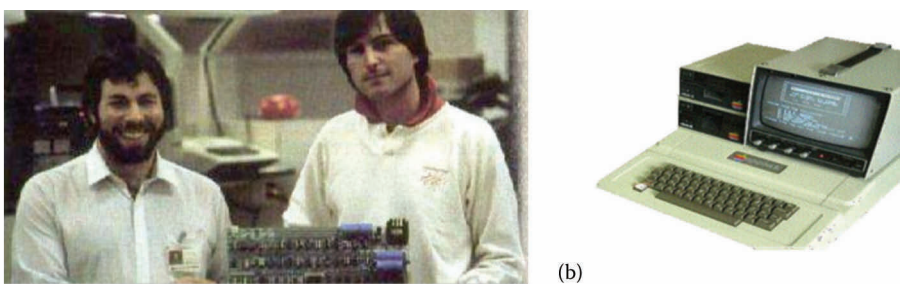


FIGURE 1.14 (a) Jobs and Wozniak with the board for Apple I and (b) the Apple II.

Inc. When it went public in 1980, its stock value was $117 million; three years later it was worth $985 million.

Vacuum tubes coexisted with their progeny, the transistor, and even with ICs for a short while. Although solid-state technology overwhelmingly dominates today's world of electronics, vacuum tubes are still holding out in some areas. You might, for example, still have a CRT (cathode ray tube) as your television or computer screen. Tubes also remain in two small but vibrant areas for entirely different reasons. The first involves microwave technology, which still relies on vacuum tubes for their power-handling capability at high frequencies. The other—the creation and reproduction of music—is a more complicated story. Tubes distort signals differently than transistors when overdriven, and this distortion is regarded as being more "pleasant" by much of the music community.

Extrapolating back to 1961, Gordon Moore in 1965 (Figure 1.15), while at Fairchild, predicted that transistors would continue to shrink, doubling in density on an IC every 18–24 months, while the price would continue to come down—this prediction we know today as Moore's Law. History has proven Moore right as evidenced by past and projected feature sizes of ICs in the International Technology Roadmap for Semiconductors (ITRS) shown in Appendix 1A and on the Internet at http://public.itrs.net (updated in 2007).[5] In this International Technology Roadmap for Semiconductors, technology modes have been defined. These modes are the feature sizes that have to be in volume manufacturing at a fixed date (year of production). The feature size is defined as half-pitch, i.e., half of a dense pair of lines and spaces (see figure in Appendix 1A).

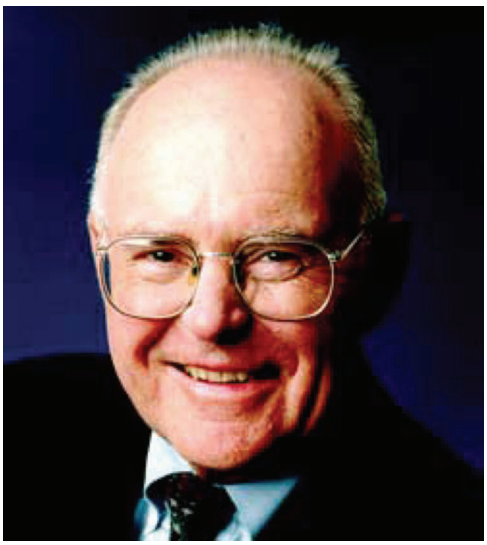


FIGURE 1.15 Gordon Moore, cofounder of Intel.

TABLE 1.1 Integration Scale and Circuit Density

IC Evolution	Acronym	Number of Logic Gates	Year of Introduction
Zero-scale integration	ZSI	1	1950
Small-scale integration	SSI	2–30	1965
Medium-scale integration	MSI	30–10^3	1970
Large-scale integration	LSI	10^3–10^5	1980
Very large-scale integration	VLSI	10^5–10^7	1985
Ultra-large-scale integration	ULSI	10^7–10^9	1990
Giga-scale integration	GSI	10^9–10^{11}	2005
Tera-scale integration	TSI	10^{11}–10^{13}	2020

In Table 1.1 the increasing numbers of devices integrated on an IC are tabulated. As we will learn in Chapters 1 and 2 on lithography in Volume II, new lithography techniques, novel device structures, and the use of new materials drive Moore's Law.

The state of the art in ICs today is a 2-GB DRAM* with 60-nm features (Samsung). Intel introduced the Core 2 Quad "Kentsfield" chip in January 2007, a chip featuring a 65-nanometer technology mode.

The 32-nm node should be achieved in 2009. The first Moore's Law is the good news, but there is a second Moore's Law that is a bit problematic; this second law states that the cost of building a chip factory doubles with every other chip generation, i.e., every 36 months. Today's technology involves Si wafers with a 12-in. diameter and factories that cost $3–4 billion to construct. With this type of start-up costs, few countries can afford to enter the IC market, and the search is on for alternative, less-expensive

* Dynamic random access memory. A type of memory component used to store information in a computer system. "Dynamic" means the DRAMs need a constant "refresh" (pulse of current through all the memory cells) to keep the stored information.

bottom-up NEMS techniques (below). From Appendix 1B, the IC market for 2003 was $166 billion worldwide, with 2004 projected at $241 billion (data by the World Semiconductor Trade Statistics [WSTS]; http://www.wsts.org).[6] Also from Appendix 1B we learn that the IC business is feeding a trillion-dollar electronic equipment business. It is worth pointing out that China is expected to control 5% of the IC market by 2010.

MEMS

Single-crystal silicon is not only an excellent electronic material but it also exhibits superior mechanical properties; the latter gave birth to the **m**icro**e**lectro**m**echanical **s**ystems (MEMS) field on the coattails of the IC industry. Originally MEMS constituted mostly mechanical types of devices based on single-crystal silicon with at least one or more of their dimensions in the micrometer range. As MEMS applications broadened, in Europe the acronym MST for ***m**icrosystem **t**echnology* became more popular. In Japan one refers to *micromachining*, and *microengineering* is popular in the United Kingdom. Development of single-crystal silicon mechanical MEMS involves the fabrication of micromechanical parts, e.g., a thin membrane in the case of a pressure sensor or a cantilever beam for an accelerometer. These micromechanical parts are fabricated by selectively etching areas of a Si substrate away to leave behind the desired geometries. The terms *MEMS* and *micromachining* came into use in 1982 to name these fabrication processes. Around the same time references to "bulk" micromachining techniques also appeared. Richard Feynman's December 26, 1959 presentation "There's Plenty of Room at the Bottom" is considered by many to be the starting bongo for MEMS (Figure 1.16) (http://www.its.caltech.edu/~feynman/plenty.html),[7] but in a practical sense, it was the invention of the transistor and the processes developed to fabricate transistors, six years earlier, that enabled MEMS.

An early milestone for the use of single-crystal silicon in MEMS was the 1956 discovery of porous Si by Uhlir.[8] His discovery eventually led to all types of interesting new, single-crystal Si-based devices, from reference electrodes for electrochemical sensors, biosensors, quantum structures, and permeable membranes to photonic crystals and photoluminescent and electroluminescent devices.

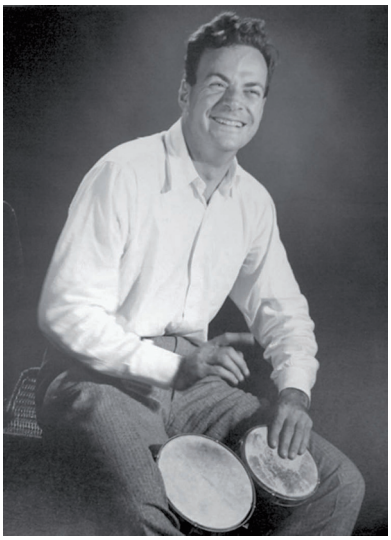


FIGURE 1.16 Richard Feynman on the bongo drums.

The first impetus for the use of single-crystal silicon as a micromechanical element in MEMS can be traced to the discovery of its large piezoresistance. Piezoresistance is the change in the resistivity of certain materials as a result of an applied mechanical strain. Charles Smith, of the Case Institute of Technology (now part of the Case Western Reserve University), during a sabbatical leave at Bell Labs in 1953, studied the piezoresistivity of semiconductors and published the first paper on the piezoresistive effect in Si and Ge in 1954.[9] The piezoresistive coefficients Smith measured demonstrated that the gauge factor* of Si and Ge strain gauges (see Figure 1.17) was 10–20 times larger than the gauge factor of metal film strain gauges, and, therefore, semiconductor gauges were expected to be much more sensitive.

Motivated by these results, companies such as Kulite and Honeywell started developing Si strain gauges commercially from 1958 onward. Pfann and colleagues, in 1961, proposed a dopant diffusion technique for the fabrication of silicon piezoresistive sensors for the measurement of stress, strain, and pressure.[10] Based on this idea, Kulite integrated

* A strain gauge is a device used to measure deformation (strain) of an object. The gauge factor of a strain gauge relates strain to change in electrical resistance.

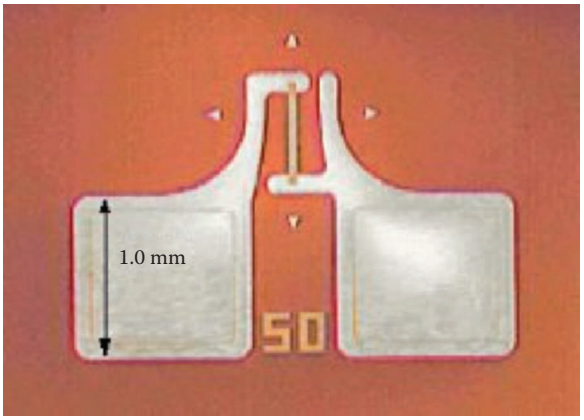


FIGURE 1.17 The Si single-crystal gauge element can be seen as the vertical bar centered between the two solder pads. The single-crystal silicon strain gauge offers sensitivities 20–50 times greater than metal foil gauges. A microphotograph of the LN-100. (BF Goodrich Advanced Micro Machines.)

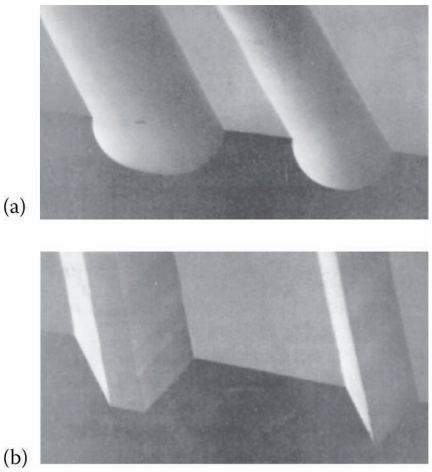


FIGURE 1.19 Isotropic and anisotropic etching profiles in single-crystal Si. Isotropic etching of grooves in (100) Si (a) and anisotropic etching of grooves in (100) Si (b) using rectangular mask openings. Features are in the 100-μm range.

Si strain gauges on a thin Si substrate with diffused resistors in 1961. As early as 1962, Tufte and coworkers at Honeywell, using a combination of wet isotropic etching, dry etching (using a plasma instead of a solution), and oxidation, made the first thin Si piezoresistive diaphragms for pressure sensors.[11] Isotropic etching of Si had been developed earlier for transistor fabrication. In the mid-1960s, Bell Labs started work on single-crystal silicon etchants with directional preferences, i.e., anisotropic etchants, such as mixtures of, at first, KOH, water, and alcohol and later KOH and water.[12] Both chemical and electrochemical anisotropic etching methods were pursued. The aspect ratio (height-to-width ratio) of features in MEMS is typically much higher than in ICs (Figure 1.18). The first high-aspect-ratio cuts in silicon were used in the fabrication of dielectrically isolated structures in ICs such as those for beam leads. In the mid-1970s, a surge of activity in anisotropic etching was associated with the work on V-groove and U-groove transistors. Isotropic and anisotropic etching profiles are compared in Figure 1.19. Figure 1.19a shows the isotropic etching of grooves in (100) Si, and Figure 1.19b shows the anisotropic etching of grooves in (100) Si. In both cases, rectangular mask openings are used.

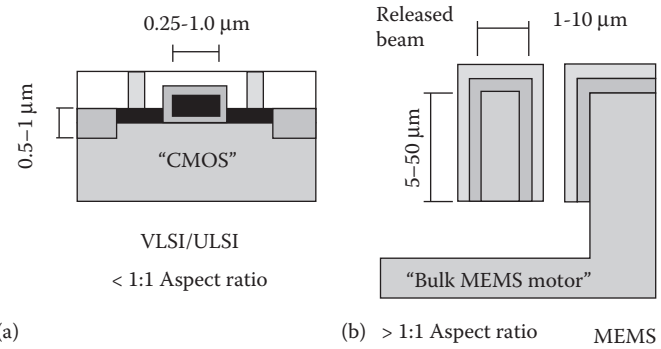


FIGURE 1.18 Aspect ratio (height-to-width ratio) typical in (a) fabrication of integrated circuits and (b) microfabricated component.

Most single-crystal silicon MEMS devices feature bonding of one Si wafer to another or to a differing substrate, say a glass pedestal, and some MEMS involve cavity-sealing techniques, perhaps for a vacuum reference or to accommodate a deflecting cantilever beam. The most prominent techniques developed to achieve these features are field-assisted bonding, invented by Wallis and Pomerantz in 1969,[13] and Si fusion bonding (SFB) by Shimbo in 1986.[14] Field-assisted thermal bonding, as shown in Figure 1.20, also known as *anodic* bonding, *electrostatic* bonding, or the *Mallory* process, is commonly used for joining glass to silicon at high temperatures (e.g., 400°C) and high voltages (e.g., 600 V). The ability to bond two Si wafers directly, at high temperatures (>800°C) in an oxidizing environment, without intermediate layers or applying an electric field, simplified the fabrication of many devices in silicon fusion bonding (SFB).

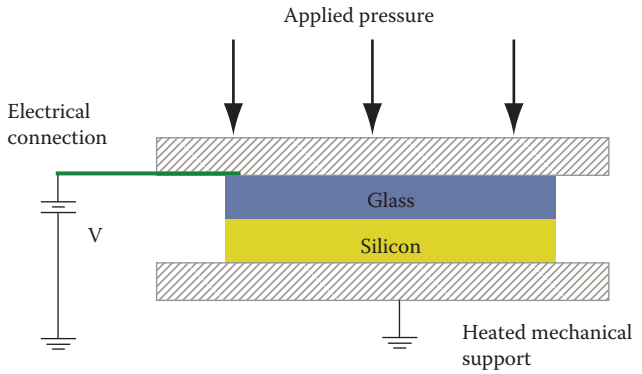


FIGURE 1.20 During anodic bonding, the negative potential applied to the borosilicate glass plate, which has been heated to 500°C, allows the migration of positive ions (mostly Na^+) away from the wafer's interface, creating a strong electric field between the glass and the Si wafer.

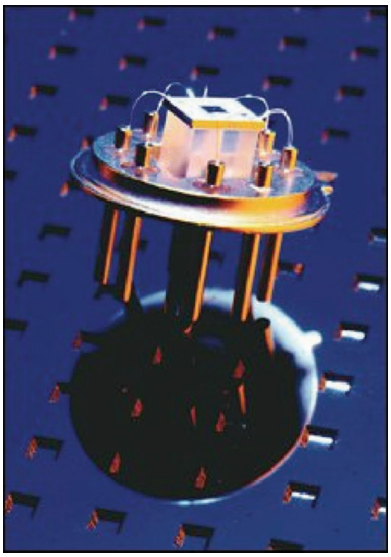


FIGURE 1.21 Piezoresistive pressure sensor featuring a Si/glass bond achieved by anodic bonding. The Bosch engine control manifold absolute pressure (MAP) sensor is used in automobile fuel injection systems. By measuring the manifold pressure, the amount of fuel injected into the engine cylinders can be calculated. Micromachined silicon piezoresistive pressure sensors are bonded at the wafer level to a glass wafer using anodic bonding before dicing. The glass pedestal that is created by this process provides stress isolation for the silicon sensor from package-induced thermal stresses. (Photo courtesy of Robert Bosch GmbH, Germany.)

The first Si accelerometer was demonstrated in 1970 at Kulite. In 1972, Sensym became the first company to make stand-alone Si sensor products. By 1974, National Semiconductor Corporation, in California, carried an extensive line of Si pressure transducers as part of the first complete silicon pressure transducer catalog.[15] Other early commercial suppliers of micromachined pressure sensor products were Foxboro/ICT, Endevco, Kulite, and Honeywell's Microswitch. To achieve better sensitivity and stability than possible with piezoresistive pressure sensors, capacitive pressure sensors were first developed and demonstrated by Dr. James Angell at Stanford University around 1977.[16] In Figure 1.21 we show a typical piezoresistive single-crystal silicon pressure sensor, with the silicon sensor anodically bonded to a glass substrate.

In many cases, it is desirable to stop the etching process when a certain cavity depth or a certain membrane thickness is reached. High-resolution silicon micromachining relies on the availability of effective etch-stop layers rather than the use of a stopwatch to control the etch depth. It was the discovery/development of impurity-based etch stops in silicon that allowed micromachining to become a high-yield commercial production process. The most widely used etch-stop technique is based on the fact that anisotropic etchants do not attack heavily boron-doped (p++) silicon layers. Selective p++ doping is typically implemented using gaseous or solid boron diffusion sources with a mask (such as silicon dioxide). The boron etch-stop effect was first noticed by Greenwood in 1969,[17] and Bohg in 1971[18] found that an impurity concentration of about $7 \times 10^{19}/cm^3$ resulted in the anisotropic etch rate of Si decreasing sharply.

Innovative, micromachined structures, different from the now mundane pressure sensors, accelerometers, and strain gauges, began to be explored by the mid- to late 1970s. Texas Instruments produced a thermal print head in 1977,[19] and IBM produced inkjet nozzle arrays the same year.[20] In 1980, Hewlett Packard made thermally isolated diode detectors,[21] and fiberoptic alignment structures were manufactured at Western Electric. Chemists worldwide took notice when Terry, Jerman, and Angell, from Stanford University, integrated a gas chromatograph on a Si wafer in 1979 as shown in Figure 1.22.[22,23] This first analytical chemistry application would eventually lead to the concept of total analytical systems on a chip, or μ-TAS. An important milestone in the MEMS world was the founding of NovaSensor,

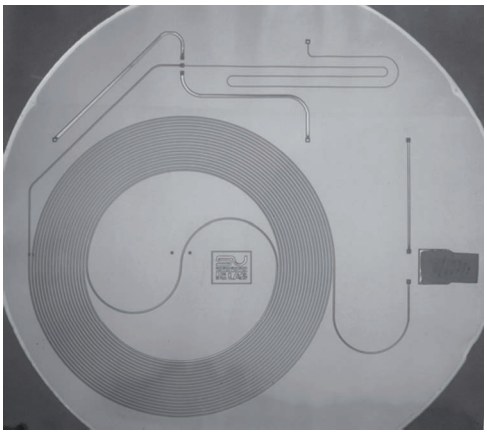


FIGURE 1.22 Gas chromatograph on a Si wafer. (Courtesy Hall Jerman.)

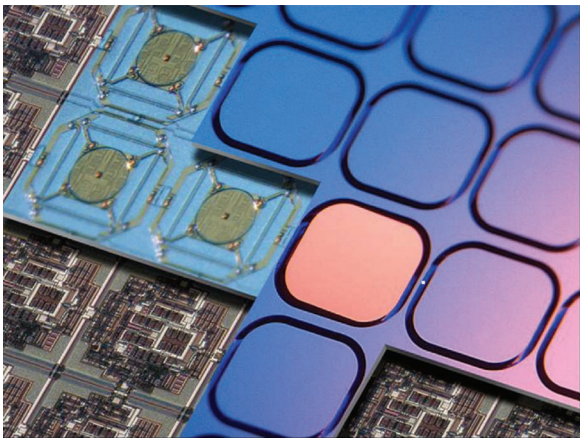


FIGURE 1.24 Integrated photonic mirror array from Transparent Networks. MEMS-VLSI integration achieved through wafer bonding. There are 1200 3D mirrors on the chip; each is 1 × 1 mm, with a ±10° tilt in two axes. (Courtesy of Janusz Bryzek.)

in 1985, by Kurt Petersen, Janusz Bryzek, and Joe Mallon (Figure 1.23). This was the first company totally dedicated to the design, manufacture, and marketing of MEMS sensors.

Kurt Petersen developed the first torsional, scanning micromirror in 1980 at IBM.[24] A more recent version of a movable mirror array is shown in Figure 1.24. Mirror arrays of this type led to the infamous stock market optical MEMS bubble of 2000, one of the bigger disappointments befalling the MEMS community.

The first disposable blood pressure transducer became available in 1982 from Foxboro/ICT, Honeywell for $40. Active on-chip signal conditioning also came of age around 1982. European and Japanese companies followed the U.S. lead more than a decade later; for example, Druck Ltd., in the United Kingdom, started exploiting Greenwood's micromachined pressure sensor in the mid-1980s. Petersen's 1982 paper extolling the excellent mechanical properties of single-crystalline silicon helped galvanize academia's involvement in Si micromachining in a major way.[25]

FIGURE 1.23 NovaSensor founders in 2003 at the Boston Transducer Meeting. Left to right: Joe Mallon, Kurt Petersen, and Janusz Bryzek.

In MEMS the need sometimes arises to build structures on both sides of a Si wafer; in this case, a double-sided alignment system is required. These systems started proliferating after the EV Group (formally known as Electronic Visions) created the world's first double-side mask aligner with bottom side microscope in 1985 (http://www.evgroup.com). In the mid-1990s, new high-density plasma etching equipment became available, enabling directional deep dry reactive ion etching (DRIE) of silicon. Dry plasma etching was now as fast as wet anisotropic etching, and as a consequence the MEMS field underwent a growth spurt.

U.S. government agencies started large MEMS programs beginning in 1993. Older MEMS researchers remember the idealistic and inspired leadership of Dr. Kaigham (Ken) Gabriel (Figure 1.25) at Defense Advanced Research Projects Agency (DARPA). Gabriel got many important new MEMS products launched.

When the first polysilicon MEMS devices, made in a process called surface micromachining pioneered at University of California, Berkeley by Muller

FIGURE 1.25 Dr. Kaigham Gabriel: Early champion of MEMS work at DARPA.

and Howe, appeared in 1983,[26] bulk single-crystal Si micromachining started to get some stiff competition. This was exacerbated when, from the mid-1990s onward, MEMS applications became biomedical and biotechnology oriented. The latter applications may involve inexpensive disposables or implants, and Si is not a preferred inexpensive substrate nor is it biocompatible. Glass and polymers became very important substrates in microfluidics, and many researchers started using a flexible rubber (polydimethylsiloxane or PDMS) as a building material in a process called soft lithography. The latter manufacturing method, which dramatically shortened the time between novel fluidic designs and their testing, was invented in the late 1990s by Harvard's Whitesides.[27]

In the early years of MEMS, it was often projected that the overall MEMS market would grow larger than that for ICs. This notion was based on the expectation of many more applications for the former than the latter. This market projection has not been fulfilled. Including nonsilicon devices such as read-write heads, one can claim a MEMS market of about 10% of the total IC market today. From Appendix 1B we learn that Si sensors and actuators amount to only 4% of IC sales.

MEMS never really constituted a paradigm shift away from the IC concept but rather a broadening of it: incorporating more diverse materials, higher aspect ratio structures, and a wider variety of uses in smaller and more fragmented applications. In almost all respects MEMS remained IC's poor cousin: using second-hand IC equipment, with less than 5% of IC sales, and no Nobel laureates to trumpet breakthrough new concepts. Today, the Si MEMS market prospects are looking much better as MEMS are finally penetrating mass consumer products from projectors, to game controllers, to portable computers to cameras, mobile phones, and iPods with MEMS digital micromirror devices (DMDs), oscillators, accelerometers, gyros, etc. Even MEMS foundries are now thriving inside and outside the United States. This new generation of MEMS products fits high-throughput production lines on large Si substrates and is succeeding in the marketplace. The IC world has started to absorb the Si-MEMS world.

Over the years, the many MEMS applications did lead to a plethora of MEMS acronyms, some of them perhaps coined by assistant professors trying to get tenure faster. Here are some attempts at 15 minutes of fame:

- BioMEMS = MEMS applied to the medical and biotechnology field
- Optical MEMS = mechanical objects + optical sources/detectors
- Power-MEMS
- C-MEMS (carbon MEMS for this author but ceramic MEMS for others)
- HI-MEMS = hybrid insect-microelectromechanical systems
- RF-MEMS = radiofrequency MEMS
- Cif-MEMS = CMOS IC Foundry MEMS
- COTS MEMS = commercial off-the-shelf MEMS
- MOEMS = microoptical electromechanical systems
- P-MEMS = polymer MEMS
- CEMS = cellular engineering microsystems
- HARMEMS = high-aspect-ratio MEMS

We do expect that there are many more MEMS applications yet to be realized and that MEMS will facilitate the handshake between the macro and nano world in nanoelectromechanical systems (NEMS).

In Table 1.2 we sketch our attempt at a Si/MEMS history line. Many more MEMS milestones than listed in the preceding text are captured here.

Table 1.2 MEMS History

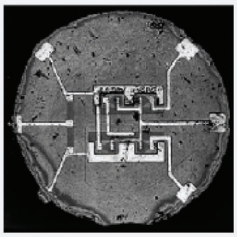
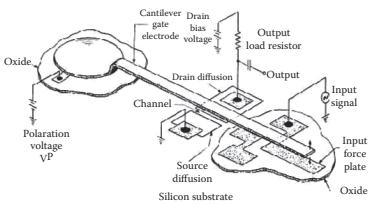
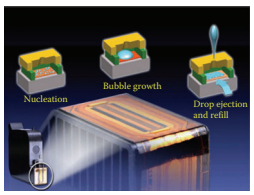
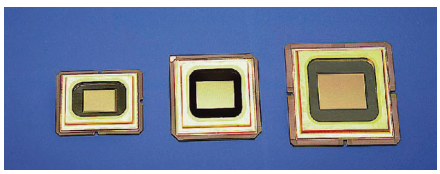
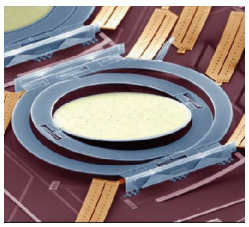
Year	Fact
1824	Berzelius discovers Si
1910	First patent on the MOS transistor concept
1927	Field effect transistor patented (Lilienfield)
1939	First pn junction transistor (J. Bardeen, W.H. Brattain, W. Shockley)
1947 (23 December) 	Invention of the transistor made from germanium at Bell Telephone Laboratories
1954	Evidence of piezoresistive effect in Si and Ge by Smith[9]
1956	An early milestone for the use of single-crystal silicon in MEMS was the discovery of porous Si by Uhlir[8]
1958  1960s IC (1st Monolithic BJT IC, 4BJT)	Jack Kilby of Texas Instruments invents the IC, using GE devices. A patent was issued to Kilby in 1959. A few months later, Robert Noyce of Fairchild Semiconductor announced the development of a planar Si IC
1958	Silicon strain gauges commercially available
1958	First IC (oscillator)
1959	R. Feynman famous talk: "There's Plenty of Room at the Bottom"[7]
1961	Fabrication of the first piezoresistive sensor, pressure (Kulite)
1967	Anisotropic deep silicon etching (H.A. Waggener[12])
1967 	First surface micromachining process (H. Nathanson[28]): resonant gate before it was called MEMS
1969–1970	Anodic bonding of glass to Si[13]
1972	National Semiconductor: commercialize a Si MEMS pressure sensor
1975	Gas chromatograph on a Si wafer by S.C. Terry, J.H. Jerman, and J.B. Angell at Stanford University[23]
1977	First capacitive pressure sensor (Stanford)[16]
1977 	IBM–HP: micromachined ink-jet nozzle[24]
1980	K.E. Petersen, silicon torsional scanning mirror[24]

Table 1.2 MEMS History (Continued)

Year	Fact
1982	Review paper "Silicon as a mechanical material" published by K.E. Petersen[25]
1982	Disposable blood pressure transducer (Foxboro/ICT, Honeywell, $40)
1982	The use of x-ray lithography in combination with electroplating and molding (or LIGA), introduced by Ehrfeld and his colleagues[29]
1983	Integrated pressure sensor (Honeywell)
1983	"First" polysilicon MEMS device (Howe, Muller UCB[26]); see also Nathanson in 1967[28]
1986	Silicon to silicon wafer bonding (M. Shimbo[14])
1987	Texas Instrument's Larry Hornbeck invents the digital micromirror devices (DMDs)
1988	Rotary electrostatic side drive motors (Fan, Tai, Muller[30]) Electrostatic micromotor (UC-Berkeley BSAC)
1988	First MEMS conference (first transducers conference held in 1987)
1989 	Lateral comb drive (Tang, Nguyen, Howe[31])
1990	The concept of miniaturized total chemical analysis system or μ-TAS is introduced by Manz et al.[32] This may be seen as the beginning of BIOMEMS
1992	Grating light modulator (Solgaard, Sandejas, Bloom)
1992	First MUMPS process (MCNC) (with support of DARPA). Now owned by MEMSCAP
1992	First MEMS CAD tools: MIT, S.D. Senturia, MEMCAD 1.0 Michigan, Selden Crary, CAEMEMS 1.0
1992	Single-crystal reactive etching and metallization (SCREAM) developed at Process (Cornell)
1993	Analog devices: commercialize multiaxis accelerometer integrating electronics (ADXL50)
1995	Intellisense Inc. introduces MEMS CAD IntelliSuite. MEMCAD 2.0 is launched, and ISE introduces SOLIDIS and ICMAT
1996  TI's VGA (640 × 480), the SVGA (800 × 600), and the XGA (1024 × 768)	The first digital mirror device (DMD)–based products (Texas Instruments) appear on the market
1996	DRIE (Bosch Process)
1997	Printing meets lithography when George M. Whitesides et al. at Harvard discover soft lithography[27]
1998	First PCR-microchips
1998	Sandia's ultraplanar multilevel technology SUMMiT-IV and -V technologies. Four- and five-level poly-Si processes

(*continued*)

Table 1.2 MEMS History (Continued)

Year	Fact
1999  Affymetrix genechip	DNA microarray techniques
1999	Electrokinetic platforms (Caliper, Aclara, and Agilent)
2000	Nortel buys Xros for $3.25 billion
2002  Lucent 3D optical switch	The telecom recession puts many things on standby
2004	MEMS rebuilds. First application of accelerometer in consumer electronics (CE) to hard drive protection in notebooks. IBM puts dual-axis accelerometer in the notebook (now Lenovo)
2006	Sony (PS3) and Nintendo (Wii) introduce motion-based game controllers
2007	Apple announces the iPhone with motion-based features

NEMS

The criteria we use in this book for classifying something as a nanoelectromechanical system (NEMS) are not only that the miniaturized structures have at least one dimension that is smaller than 100 nanometers, but also that they are crafted with a novel technique (so beer making is out) or have been intentionally designed with a specific nanofeature in mind (so medieval church stained glass is out). This definition fits well within the one adopted by the National Nanotechnology Institute (NNI; http://www.nano.gov/html/facts/whatisnano.html):

1. Nanotechnology involves R&D at the 1- to 100-nm range.
2. Nanotechnology creates and uses structures that have novel size-based properties.
3. Nanotechnology takes advantage of the ability to control or manipulate at the atomic scale.

Paul Davis (http://cosmos.asu.edu), a theoretical physicist and well-known science popularizer, said, "The nineteenth century was known as the machine age, the twentieth century will go down in history as the information age and I believe the twenty-first century will be the quantum age." We believe that the current nanotechnology revolution, underpinned by quantum mechanics, is already leading the way toward that reality.

The manufacture of devices with dimensions between 1 and 100 nanometers is either based on top-down manufacturing methods (starting from bigger building blocks, say a whole Si wafer, and chiseling them into smaller and smaller pieces by cutting, etching, and slicing), or it is based on bottom-up manufacturing methods (in which small particles such as atoms, molecules, and atom clusters are added for the construction of bigger functional constructs). The top-down approach to nanotechnology we call *nanofabrication* or *nanomachining*, an extension of

the MEMS approach. The bottom-up approach we like to refer to as *nanochemistry*. An example of this second approach is the self-assembly of a monolayer (SAM) from individual molecules on a gold surface. Bottom-up methods are nature's way of growing materials and organisms, and in biomimetics one studies how nature, through eons of time, developed manufacturing methods, materials, structures, and intelligence and tries to mimic or replicate what nature does in the laboratory to produce MEMS or NEMS structures.

A history-line with the most important NEMS milestones on it is difficult to put together as so many authors of such charts automatically include themselves or their institution on it first (one author of an early MEMS/NEMS timeline puts himself on it three times and his institution four times!). It sometimes seems that science and engineering are starting to resemble FOX News more by the day. What follows are some milestones toward nanotechnology that many scientists/engineers might agree to.

Norio Taniguchi introduced the term *nanotechnology* in 1974, in the context of traditional machining with tolerances below 1 micron. The 1959 Feynman lecture "There's Plenty of Room at the Bottom," which helped launch the MEMS field (see above), was geared more toward NEMS than MEMS (http://www.its.caltech.edu/~feynman/plenty.html).[7] Feynman proclaimed that he knew of no principles of physics that would prevent the direct manipulating of individual atoms. In his top-down *gedanken* experiment, he envisioned a series of machines each an exact duplicate, only smaller and smaller, with the smallest in the series being able to manipulate individual atoms (see Figure 1.26).

In 1981, Gerd Binning and Heinrich Rohrer of IBM Zurich invented the scanning tunneling microscope (STM), enabling scientists to see and move individual atoms. Such a microscope, shown in Figure 1.27, measures the amount of electrical current flowing between a scanning tip and the conductive surface that is being measured. This unexpectedly simple instrument allowed for the imaging of micro- and nanostructures, catapulted nanotechnology onto the world stage, and got its inventors the 1986 Nobel Prize (http://www.zurich.ibm.com/imagegallery/st/nobelprizes). Just as 350 years before the microscope changed the way we

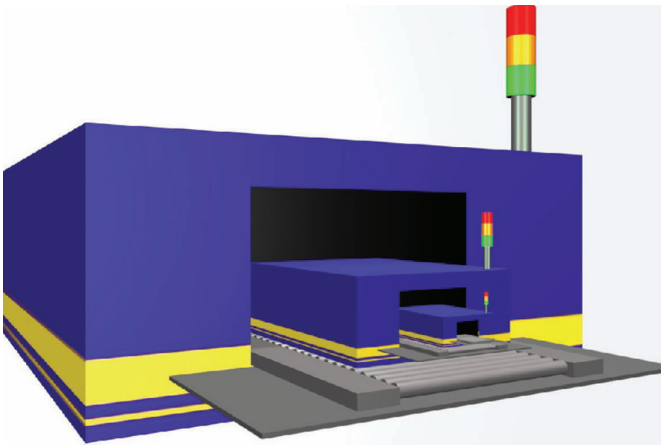


FIGURE 1.26 Master and slave hands on a set of Feynman machines.

viewed the world, the STM impacts our current view of biology, chemistry, and physics.

In fast succession, a series of similar instruments, all called scanning proximal probes, followed the introduction of the STM. For example, Binnig, Quate (Stanford), and Gerber (IBM) developed the atomic force microscope (AFM) in 1986. An AFM, in contact mode, measures the repulsive force interaction between the electron clouds on the probe tip atoms and those on the sample—making it possible to image both insulating and conducting surfaces. This results in the visualization of the interactions

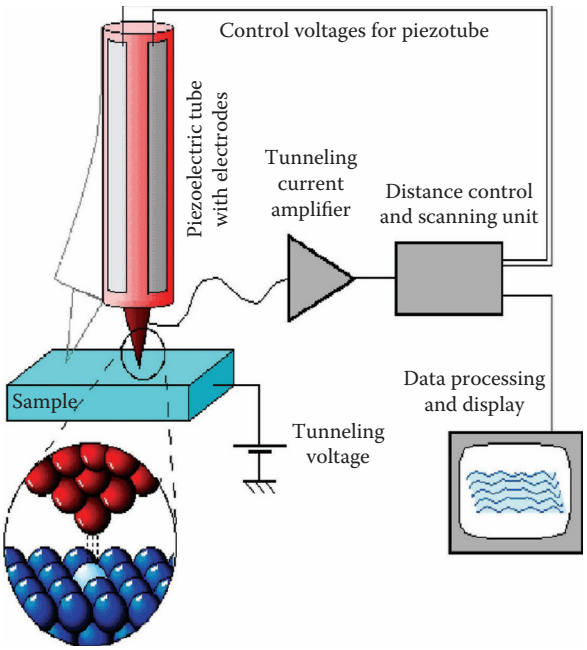


FIGURE 1.27 Scanning tunneling microscope (STM). Operational principle of an STM. (Courtesy Michael Schmidt, TU Wien.)

between molecules at the nanoscale, thus increasing our ability to better understand the mechanism of molecular and biological processes. Other forms of scanning probe microscopes—those that do not depend on tunneling or forces between a probe tip and a sample surface—have also been demonstrated. Examples include the scanning thermal microscope, which responds to local thermal properties of surfaces, the scanning capacitance microscope for dopant profiling, and the near-field scanning optical microscope (NSOM, also known as SNOM). In the latter instrument,[33] the wavelength limitation of the usual far-field optics of a light microscope is avoided by mounting the light detector (say an optical fiber) on an AFM tip, at a distance from the sample that is a fraction of the wavelength used; this way it is possible to increase the resolution of a light microscope considerably. These new tools were an important catalyst behind the surge in nanotechnology activities worldwide, and this illustrates that progress in science is inextricably linked to the development of new measurement tools.

The discovery, in the early 1980s at Bell Labs by David L. Allara (now at Pennsylvania State University) and Ralph G. Nuzzo (now at University of Illinois, Urbana-Champaign) of the self-assembly of disulfide and, soon thereafter, of alkanethiol monolayers (SAMs) on metal surfaces coincided with the maturation of STM technology.[34,35] SAMs, especially on Au, turned out to be a valuable type of sample for STM investigation, showing these films to spontaneously assemble into stable and highly organized molecular layers, bonding with the sulfur atoms onto the gold and resulting in a new surface with properties determined by the alkane head group.

Like SAMs, dendrimers, which are branching polymers sprouting successive generation of branches off like a tree, are an important tool for bottom-up nanotechnologists. Dendrimers (from the word *dendron*, Greek for tree) were invented, named, and patented by Dr. Donald Tomalia (now CTO at Dentritic NanoTechnologies, Inc.) in 1980 while at Dow Chemical.[36]

In 1970, Arthur Ashkin was the first to report on the detection of optical scattering and gradient forces on micron-sized particles.[37] In 1986, Ashkin and colleagues reported the first observation of what is now commonly referred to as an optical trap, i.e., a tightly focused beam of light capable of holding microscopic particles stable in three dimensions.[38] One of the authors of this seminal 1986 paper, Steven Chu (Figure 1.28), would go on to make use of optical tweezing techniques in his work on cooling and trapping of atoms. Where Ashkin was able to trap larger particles (10 to 10,000 nanometers in diameter), Chu extended these techniques to the trapping of individual atoms (0.1 nanometer in diameter). This research earned him the 1997 Nobel Prize in Physics (with Claude-Cohen Tannoudji and William D. Phillips). In another heralded experiment, Steven Chu was also the one who demonstrated that by attaching polystyrene beads to the ends of DNA one can pull on the beads with laser tweezers to stretch the DNA molecule (http://www.stanford.edu/group/chugroup/steve_personal.html).

FIGURE 1.28 Steven Chu (Stanford University), recipient of the 1997 Nobel Prize in Physics for his work on cooling and trapping of atoms.

In 1985, Robert F. Curl Jr., Harold W. Kroto, and the late Richard E. Smalley serendipitously (while investigating the outer atmosphere of stars) discovered a new form of carbon: buckminsterfullerene, also known as buckyball or C60, shown in Figure 1.29.[39] They were awarded the Nobel Prize in 1996.

Perhaps a more important discovery, because of its generality and broader applicability, is the one by NEC's Sumio Iijima, who, in 1991, discovered

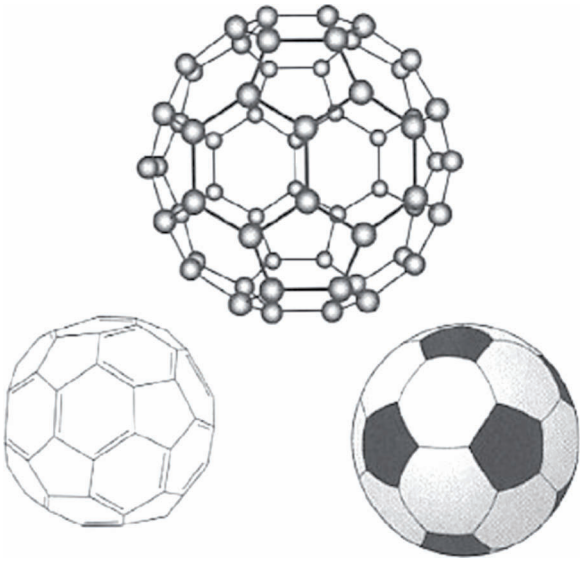


FIGURE 1.29 Buckminsterfullerene C60 or buckyball with 60 atoms of carbon; each is bound to three other carbons in an alternating arrangement of pentagons and hexagons.

carbon nanotubes, with an electrical conductivity that is up to six orders of magnitude higher than that of copper (http://www.nec.co.jp/rd/Eng/innovative/E1/myself.html). Like buckyballs, cylindrical nanotubes each constitute a lattice of carbon atoms, and each atom is again covalently bonded to three other carbons.

Carbon nanotubes exist as single-walled (SWNT) and multiwalled (MWNT); the ones depicted in Figure 1.30a are multiwall nanotubes. Unique among the elements, carbon can bond to itself to form extremely strong two-dimensional sheets, as it does in graphite, as well as buckyballs and nanotubes.

Cees Dekker demonstrated the first carbon nanotube transistor in 1998 at the Delft University of Technology[40] (see Figure 1.30b). In this device, a semiconducting carbon nanotube of only about 1 nm in diameter bridges two closely separated metal electrodes (400 nm apart) atop a silicon surface coated with silicon dioxide. Applying an electric field to the silicon (via a gate electrode) turns on and off the flow of current across the nanotube by controlling the movement of charge carriers in it. By carefully controlling the formation of metal gate electrodes, Dekker's group (http://www.ceesdekker.net) was able to create transistors with an output signal 10 times stronger than the input. At around the same time, the first nanotransistor was built at Lucent Technologies (1997). The MOS semiconductor transistor was 60 nm wide, including the source, drain, and gate; its thickness was only 1.2 nm. Other companies have since built smaller nanotransistors.

At one point, in the late 1990s, it was—with just a bit of exaggeration—hard to find a proposal to a government agency that did not involve carbon nanotubes. But perhaps more real progress was mapped in the meantime in the area of nanocrystals or quantum dots (QD).

Quantum dots possess atom-like energy states. The behavior of such small particles was beginning to be understood with work by the Russians Ekimov and Efros from 1980 to 1982.[41,42] They recognized that nanometer-sized particles of CdSe, with their very high surface-to-volume ratio, could

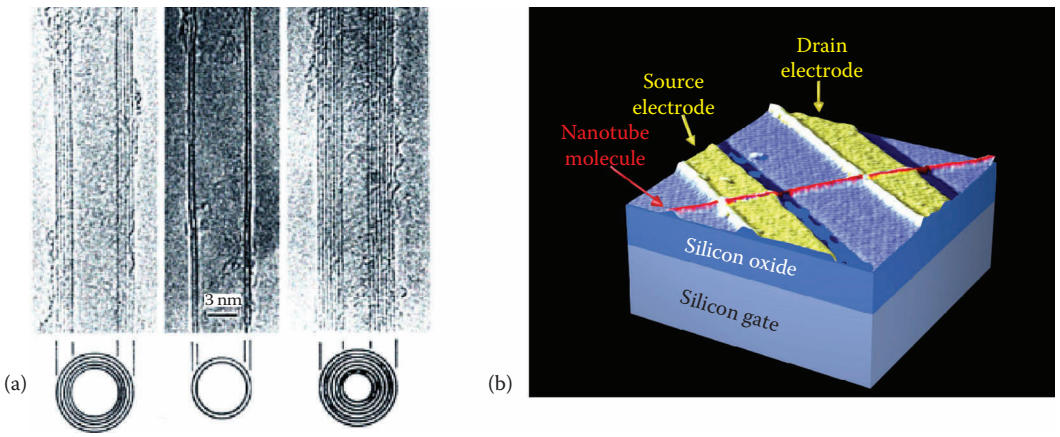


FIGURE 1.30 (a) Multiwall nanotubes with Russian inset doll structure; several inner shells are shown a typical radius of the outermost shell >10 nm. (From S. Iijima, *Nature* 354, 54, 1991.) (b) First nanotube transistor. This three-terminal device consists of an individual semiconducting nanotube on two metal nanoelectrodes with the substrate as a gate electrode.

trap electrons and that these trapped electrons might affect the crystal's response to electromagnetic fields, that is, absorption, reflection, refraction, and emission of light. Louis E. Brus, a physical chemist at Bell Laboratories at the time, and now at Columbia University, put this to practice when he learned to grow CdSe nanocrystals in a controlled manner.[43,44] Murray, Norris, and Bawendi synthesized the first high-quality quantum dots in 1993.[45] Crystallites from –12 to –115 Å in diameter with consistent crystal structure, surface derivatization, and a high degree of monodispersity were prepared in a single reaction based on the pyrolysis of organometallic reagents by injection into a hot coordinating solvent. The confinement of the wave-functions in a nanocrystal or quantum dot lead to a blue energy shift, and by varying the particle size one can produce any color in the visible spectrum, from deep (almost infra-) reds to screaming (almost ultra-) violet as illustrated in Figure 1.31. Today, quantum dots form an important alternative to organic dye molecules. Unlike fluorescent dyes, which tend to decompose and lose their ability to fluoresce, quantum dots maintain their integrity, withstanding many more cycles of excitation and light emission (they do not bleach as easily!). Combining a number of quantum dots in a bead conjugated to a biomolecule is used as a spectroscopic signature—like a barcode on a commercial product—for tagging those biomolecules.

Carbon nanotubes are only one type of nanowire. In terms of investigating and exploiting quantum confinement effects, semiconductor wires, with diameters in the 10s of nanometers, often single crystalline, represent the smallest dimension for efficient transport of electrons and excitons and are the logical interconnects and critical devices for nanoelectronics and nanooptoelectronics of the future. Over the past decade, there has been major progress in chemical synthesis technologies for growing these nanoscale semiconductor wires. As originally proposed by R.S. Wagner and W.C. Ellis from Bell Labs for the Au-catalyzed Si whisker growth, a vapor-liquid-solid mechanism is still mostly used.[46] But the field got a shot in the arm (a rebirth so to speak) with efforts by Charles Lieber (Harvard), Peidong Yang (http://www.cchem.berkeley.edu/pdygrp/main.html), James Heath (http://www.its.caltech.edu/~heathgrp), and Hongkun Park (http://www.people.fas.harvard.edu/~hpark). Lieber's group at Harvard (http://cmliris.harvard.edu) reported arranging indium phosphide semiconducting nanowires into a simple configuration that resembled the lines in a tick-tack-toe board. The team used electron beam lithography to place electrical contacts at the ends of the nanowires to show that the array was electronically active. The tiny arrangement was not a circuit yet, but it was the first step, showing that separate nanowires could communicate with one another.

Molecules are 30,000 times smaller than a transistor (180 nm on a side), so obviously it is of some use to investigate whether molecules can act as switches. Mark Ratner and Ari Aviram had suggested this as far back as 1974.[47] The suggestion remained a pipe dream until the advent of scanning probe microscopes in the 1980s, which gave researchers finally

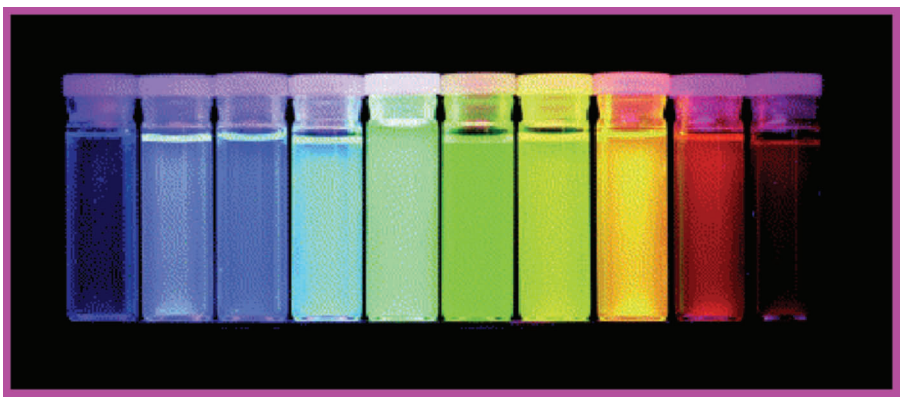


FIGURE 1.31 Different-sized quantum dots in response to near-UV light. Also composition of core affects wavelength. Red: bigger dots! Blue: smaller dots!

the tools to probe and move individual molecules around. This led to a large number of studies in the late 1990s that demonstrated that individual molecules can conduct electricity just like metal wires, and turning individual molecules into switches came not far behind. In 1997, groups led by Robert Metzger (http://bama.ua.edu/~rmmgroup) of the University of Alabama, Tuscaloosa, and Mark Reed of Yale University (http://www.eng.yale.edu/reedlab) created molecular diodes. In July 1999, another group headed by James Heath and Fraser Stoddart of the University of California, Los Angeles (UCLA) (http://stoddart.chem.ucla.edu) also created a rudimentary molecular switch, a molecular structure that carries current but, when hit with the right voltage, alters its molecular shape and stops conducting. Heath's team placed molecules called rotaxanes, which function as molecular switches, at each junction of a circuit. By controlling the input voltages the scientists showed that they could make 16-bit memory circuits work. The field of moletronics was born.

As shown in Figure 1.32, rotaxanes are "mechanically linked" molecules that consist of a dumbbell-shaped molecule, with a cyclic molecule linked around it between the two ends. The two ends of the dumbbell molecule are very big and prevent the cyclic molecule from slipping off the end. A number of factors (e.g., charge, light, pH) can influence the position of the cyclic molecule on the dumbbell.

The use of x-ray lithography in combination with electroplating and molding (or LIGA), introduced by Ehrfeld and his colleagues in 1982,[29] demonstrated to the world that lithography may be merged with more traditional manufacturing processes to make master molds of unprecedented aspect ratios and tolerances to replicate microstructures in ceramics, plastics, and metals. The hard x-rays used enable nano-sized patterns to be printed.

At around 1997, Whitesides et al. introduced soft lithography, including the use of pattern transfer of self-assembled monolayers (SAMs) by elastomeric stamping.[27] This technique formed a bridge between top-down and bottom-up machining; a master mold is made based on "traditional" lithography, and the stamp generated from this master is inked with SAMs to print (stamp) substrates with nano-sized patterns.

Imposing boundaries on photons, by making them move in a material with a periodic dielectric constant in one, two, or three directions, leads to photonic crystals. Photonic crystals were first studied by Lord Rayleigh in 1887, in connection with the peculiar reflective properties of a crystalline mineral with periodic "twinning" planes.[48] He identified a narrow bandgap prohibiting light propagation through the planes. This bandgap was angle-dependent because of the differing periodicities experienced by light propagating at non-normal incidences, producing a reflected color that varies sharply with angle. A similar effect is seen in nature, such as in butterfly wings (Figure 1.33) and abalone shells.

A one-dimensional periodic structure, such as a multilayer film (a Bragg mirror), is the simplest type of photonic crystal, and Lord Rayleigh showed that any such one-dimensional system has a bandgap. The possibility of two- and three-dimensionally periodic crystals with corresponding two- and three-dimensional bandgaps was suggested 100

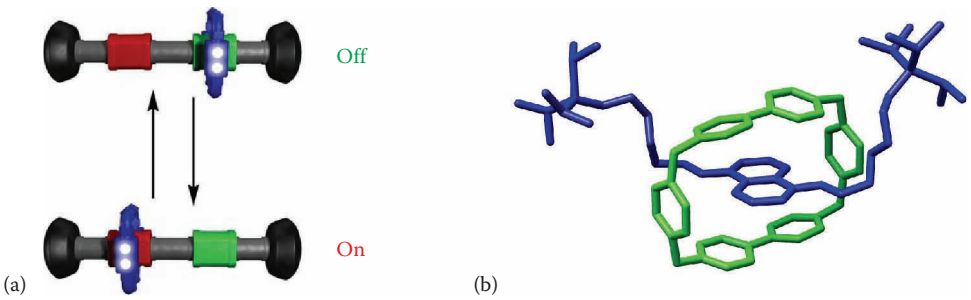


FIGURE 1.32 (a) A diagram of rotaxane. The usefulness of rotaxanes is because there are a number of positions along the dumbbell molecule that the cyclic molecule can attach to temporarily. The dumbbell can be thought of as a train track, with the positions on the dumbbell molecule as stations and the cyclic molecule as the train. (b) Crystal structure of rotaxane with a cyclobis(paraquat-p-phenylene) macrocycle.

FIGURE 1.33 A full-grown *Morpho rhetenor* butterfly, a native to South America. (From University of Southhampton. Color by nanostructure instead of dyes.)

years after Rayleigh by Eli Yablonovitch[49] (http://www.ee.ucla.edu/labs/photon/homepage.html) and Sajeev John[50] (http://www.physics.utoronto.ca/~john) in 1987. Yablonovitch (in 1991)[51] demonstrated the first microwave photonic bandgap (PBG) structure experimentally with 1-mm holes drilled in a dielectric material as illustrated in Figure 1.34 and known today as *yablonovite*. Since then, several research groups verified this prediction, which has ignited a worldwide rush to build tiny "chips" that control light beams instead of electron streams. In photonic crystals the repeat unit in the lattice is of the same size as the incoming wavelength, so homogeneous (effective) media theory cannot be applied. Photonic crystals feature lattice spacings ranging from the macroscopic (say 1 mm, for operating in the microwave domain-like yablonovite) to the 100s of nanometer range (to operate in the visible range). We cover photonic crystals here under NEMS, although only photonic crystals for the visible range qualify as nanotechnology. The potential applications of photonics are limitless, not only as a tool for controlling quantum optical systems but also in more efficient miniature lasers for displays and telecommunications, in solar cells, LEDs, optical fibers, nanoscopic lasers, ultrawhite pigments, radiofrequency antennas and reflectors, photonic integrated circuits, etc.

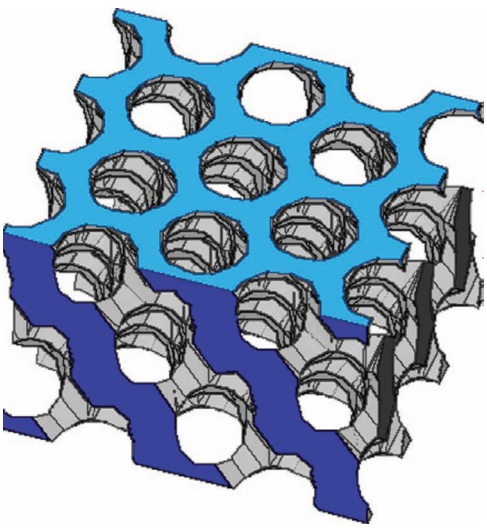


FIGURE 1.34 Holes drilled in dielectric: known now as *yablonovite*, after Yablonivitch (http://www.ee.ucla.edu/~pbmuri). The holes are 1 mm in size, and this photonic crystal is meant to operate in the microwave range.

In 1967 Victor Veselago, a Russian physicist, predicted that composite metamaterials might be engineered with negative magnetic permeability and negative permittivity.[52] Metamaterials are artificially engineered materials possessing properties that are not encountered in nature. Whereas photonic materials do exist in nature, metamaterials do not; moreover, in the case of metamaterials, the building blocks are small compared with the incoming wavelength so that effective media theory can be applied. In conventional materials the plane of the electrical field, the plane of the magnetic field, and the direction in which light travels are all oriented at right angles to each other and obey the right-hand rule. In Veselago's imaginary metamaterial, the above quantities obey a left-hand rule (as if they were reflected in a mirror). These materials would interact with their environment in exactly the opposite way from natural materials (see negative refractive index water in Figure 1.35). One intriguing prediction was that the left-hand rule would allow for a flat superlens to focus light to a point and that could image with a resolution far beyond the diffraction limit associated with far-field illumination. Veselago's prediction that such perfect lenses could be made from metamaterials lay dormant until 1996–2000, when the remarkable John Pendry, a physicist at Imperial College in London, showed that certain metals could be engineered to respond to electric fields as though the field parameters were negative.[53–58] In 2001, researchers at Imperial College and Marconi Caswell Ltd. (London) announced a magnetic resonance

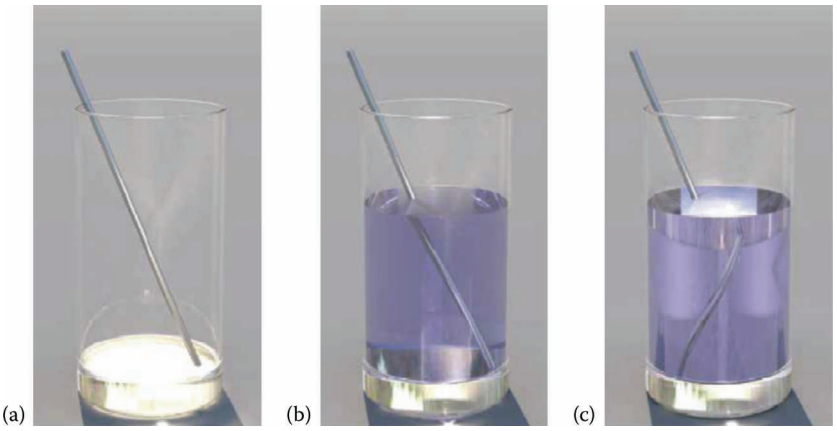


FIGURE 1.35 Refraction illustrated (a) empty glass: no refraction; (b) typical refraction with pencil in water with $n = 1.3$; (c) what would happen if the refractive index were negative with $n = -1.3$ (see metamaterials). (From Gennady Shvets, The University of Texas at Austin; http://www.ph.utexas.edu/~shvetsgr/lens.html.)

imaging system using a magnetic metamaterial based on Pendry's design.[59] Physicist Richard Shelby's group at the University of California, San Diego demonstrated a left-handed composite metamaterial that exhibited a negative index of refraction for microwaves.[60] The simple arrangement consisted of a planar pattern of copper split-ring resonators (SRRs) and wires on a thin fiberglass circuit board. Operating in the microwave range these metal patterns are large (5 mm repeat unit) but progress toward metamaterials operating in the visible was very swift. By 2005, Zhang's group at University of California, Berkeley made a 35-nm thick Ag superlens and imaged objects as small as 40 nm with 365 nm light, clearly breaking the diffraction limit of far-field imaging[61] (find Zhang's group at http://xlab.me.berkeley.edu). By 2007, a left-handed material operating in the visible range (780 nm) was demonstrated[62] by Soukoulis (http://cmpweb.ameslab.gov/personnel/soukoulis) at the U.S. Department of Energy's Ames Laboratory on the Iowa State University campus and Wegener's group from the University of Karlsruhe (http://www.aph.uni-karlsruhe.de/wegener), Germany.

As in the case of photonic crystals, only the metamaterials operating in the visible qualify as nanotechnology, but for comprehensiveness sake we cover all of them together here.

In 2000, IBM scientists placed a magnetic cobalt atom inside an elliptical coral of atoms. They observed the Kondo effect, i.e., electrons near the atom align with the atom's magnetic moment, effectively canceling it out. When the atom was placed at one focus of the elliptical coral, a second Kondo effect was observed at the other focus, even though no atom was there (see Figure 1.36). Hence some of the properties (info) carried by an atom are transferred to the other focus (www.research.ibm.com). This quantum mirage effect "reflects" information using the wave nature of the electrons rather than transmission of info using electrons in a wire. It has the potential to be able to transfer data within future nanoscale electronic circuits where wires would not work. This would allow miniaturization of circuits well below what is envisioned today.

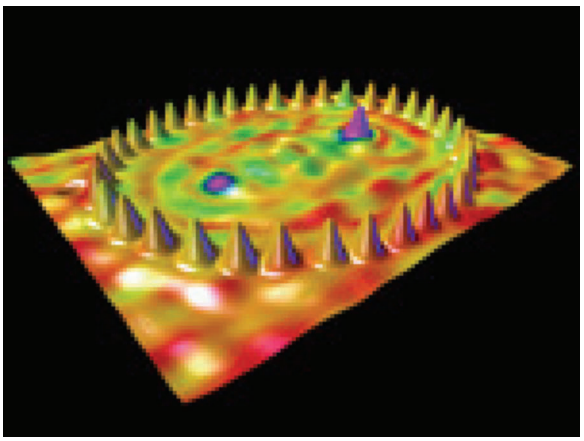


FIGURE 1.36 Quantum mirage phenomenon (http://domino.research.ibm.com/comm/pr.nsf/pages/rsc.quantummirage.html).

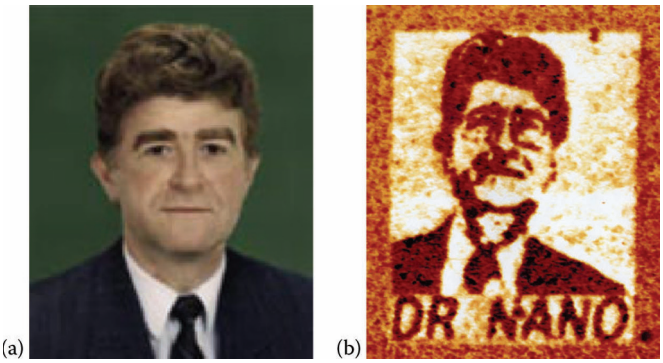


FIGURE 1.37 NSF's Dr. Mike Roco, a photo (a) and a nanograph (b). The nanograph of Dr. M. Roco was recorded at Oak Ridge National Laboratory using piezoresponse force microscopy, one of the members of the family of techniques known as scanning probe microscopy, which can image and manipulate materials on the nanoscale. Each picture element is approximately 50 nanometers in diameter; the distance from chin to eyebrow is approximately 2.5 microns. (Courtesy Dr. Roco.)

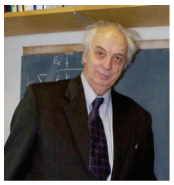
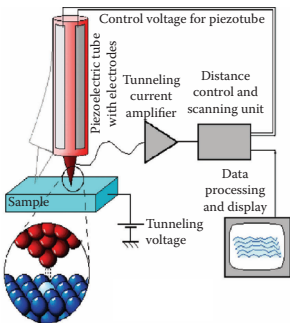
In 1999 President Clinton announced the National Nanotechnology Initiative (NNI); this first formal government program for nanotechnology accelerated the pace of nano research (the program had been around unofficially since 1996). In December 2003, George W. Bush signed the 21st Century Nanotechnology Research and Development Act. In this government NEMS program, Dr. Mike Roco (Figure 1.37) played a similar catalyzing role that Dr. Ken Gabriel played earlier in the MEMS field (see above).

Two other important nanotech promoters are Ray Kurzweil (http://www.kurzweilai.net/index.html) and K. Eric Drexel (http://www.foresight.org). Whereas Feynman continues to receive almost universal praise for his inspiring 1959 speculative talk, K. Eric Drexel, who in 1981 described mechanochemistry in his speculative paper "Molecular Engineering: An Approach to the Development of General Capabilities for Molecular Manipulation," continues to receive mostly harsh criticism—sometimes bordering on derision. In this paper and in two books,[63,64] Drexel builds nanotechnology, bottom-up, atom by atom, rather than whittling down materials as Feynman had suggested. Drexel also makes more of the fact that nature and molecular biology are proof of concept for this type of molecular technology. Drexel's early warnings about "gray goo," his emphasis on assemblers—small machines that would guide chemical bonding operations by manipulating reactive molecules—and building nanotechnology in a dry environment probably explain most of the hostility toward his work (even by those who do not even make an attempt to understand it; see for example Atkinson[65] in Nanocosm—we might as well listen to Newt Gingrich talk about nanotechnology). Drexel, unfortunately, has been associated too much with the nano pop culture. *Nano!* by Ed Regisis is an engaging and entertaining book that describes some of the researchers involved in nanotechnology; he is uncharacteristically positive about Drexler.[66]

Market projections on NEMS are today even wilder than the early ones on MEMS. Overall, though, this author believes that it is in nanotechnology, especially when considering bottom-up manufacturing, that a paradigm shift away from IC-type manufacturing is taking shape, and that it is nanotechnology that holds the potential of having a much larger impact on society than IC technology ever did. Indeed, the nanoscale is unique because at this length scale important material properties, such as electronic conductivity, hardness, or melting point, start depending on the size of the chunk of material in a way they do not at any other scale. Moreover, in biotechnology, molecular engineering already has made major progress, and the confluence of miniaturization science and molecular engineering is perhaps the most powerful new avenue for progress of humankind in general. In this regard, NEMS must be seen as a support technology to extract yet more benefits from the ongoing biotechnology revolution.

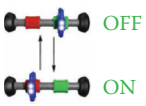
In Table 1.3 we show a milestone chart in nanotechnology. It can be argued that molecular scientists and genetic engineers were practicing nanotechnology long before the name became popular with electrical and mechanical engineers. Molecular biology or "wet nanotechnology" has been called "nanotechnology that works." But adding breakthroughs pertaining to molecular biology to this table would make it much too long. For the same reason we also did not list any IC-related milestones.

Table 1.3 A Milestone Chart in Nanotechnology

3.5 billion years ago	The first living cells emerge. Cells house nanoscale biomachines that perform such tasks as manipulating genetic material and supplying energy
400 BC	Democritus coins the word "atom," which means "not cleavable" in Greek
1857	Michael Faraday introduces "colloidal gold" to the Royal Society
1887	Photonic crystals are studied by Lord Rayleigh, in connection with the peculiar reflective properties of a crystalline mineral with periodic "twinning" planes[48]
1905	Albert Einstein publishes a paper that estimates the diameter of a sugar molecule as 1 nm. Jean-Baptist Perrin confirmed these results experimentally and was awarded the 1926 Nobel Prize for this work
1931	Max Knoll and Ernst Ruska develop the electron microscope, which enables nanometer imaging
1932	Langmuir establishes the existence of monolayers (Nobel Prize in 1932)
1959	Richard Feynman gives his famed talk "There's Plenty of Room at the Bottom," on the prospects for miniaturization[7]
1967  Victor Veselago	Victor Veselago, a Russian physicist, predicted that composite metamaterials might be engineered with negative magnetic permeability and negative permittivity[52]
1968	Alfred Y. Cho and John Arthur of Bell Laboratories and their colleagues invent molecular-beam epitaxy, a technique that can deposit single atomic layers on a surface
Early 1970s	Groups at Bell Laboratories and IBM fabricate the first two-dimensional quantum wells
1974	Norio Taniguchi conceives the word *nanotechnology* to signify machining with tolerances of less than a micron
1974	Mark Ratner and Ari Aviram suggest using molecules as switches[47]
1980	The behavior of quantum dots began to be understood with work by the Russians Ekimov and Efros in 1980–1982.[41,42] Louis E. Brus learned to grow CdSe nanocrystals in a controlled manner[43,44]
1980	Dendrimers (from the word *dendron*, Greek for tree) were invented, named, and patented by Dr. Donald Tomalia[36]
1981 	Gerd Binnig and Heinrich Rohrer create the scanning tunneling microscope, which can image individual atoms
Early 1980s	The discovery, in the early 1980s, by David L. Allara and Ralph G. Nuzzo of the self-assembly of disulfide and, soon thereafter, of alkanethiol monolayers (SAMs) on metal surfaces
1982	The use of x-ray lithography in combination with electroplating and molding (or LIGA) is introduced by Ehrfeld and his colleagues[29]
1984	Pohl develops near-field scanning optical microscope (NSOM, also known as SNOM)[33]
1985 	Robert F. Curl, Jr., Harold W. Kroto, and Richard E. Smalley discover buckminsterfullerenes, also known as buckyballs, which measure about a nanometer in diameter[39]

(*continued*)

Table 1.3 A Milestone Chart in Nanotechnology (Continued)

1986	Ashkin and colleagues reported the first observation of what is now commonly referred to as an optical trap, i.e., a tightly focused beam of light capable of holding microscopic particles stable in three dimensions[38]
1986	K. Eric Drexel publishes *Engines of Creation*, a futuristic book that popularizes nanotechnology
1987	The possibility of two- and three-dimensionally periodic crystals with corresponding two- and three-dimensional bandgaps was suggested 100 years after Rayleigh, by Eli Yablonovitch[49] and Sajeev John[50]
1989 	Donald M. Eigler of IBM writes the letters of his company's name using 35 individual xenon atoms on a nickel surface (in high vacuum and at liquid helium temperatures)
1991	Yablonovitch[52] demonstrates the first microwave photonic bandgap (PBG) structure experimentally (holes drilled in dielectric), known now as *yablonovite*
1991	Sumio Iijima of NEC in Tsukuba, Japan, discovers carbon nanotubes. The first single-walled nanotubes (SWNT) were produced in 1993
1993	The first high-quality quantum dots are synthesized by Murray, Norris, and Bawendi[45,67]
1993	Warren Robinett of the University of North Carolina and R. Stanley Williams of UCLA devise a virtual reality system connected to an STM that lets users see and touch atoms
1997	The first complete metal oxide semiconductor transistor (60 nm wide) is invented by Lucent Technologies. The key breakthrough was the 1.2-nm-thick gate oxide
1997	Whitesides et al.[27] introduced soft lithography, including the use of pattern transfer of self-assembled monolayers (SAMs) by elastomeric stamping
1998	Cees Dekker's group at the Delft University of Technology in the Netherlands creates a transistor from a carbon nanotube[40]
1999 	James Heath and Fraser Stoddart of UCLA (http://stoddart.chem.ucla.edu/) create rudimentary molecular switches, with molecules called rotaxanes, which function as molecular switches
1999	James M. Tour, now at Rice University, and Mark A. Reed of Yale University demonstrate that single molecules can act as molecular switches[68]
1999	The Clinton administration announces the National Nanotechnology Initiative, which provides a big boost in funding and gives the field greater visibility
1999 	Thermomechanical memory device, unofficially known as "Millipede," first demonstrated at IBM Zurich
2000	Eigler and other IBM scientists devise a quantum mirage—placing a magnetic atom at the focus of an elliptical ring of atoms creates a mirage atom at the other focus—transmitting info without wires
2000	John Pendry, a physicist at Imperial College in London, showed that certain metals could be engineered to respond to electric fields as though the field parameters were negative[55]
2001	Researchers at Imperial College and Marconi Caswell (London) announced a magnetic resonance imaging system using a magnetic metamaterial based on Pendry's design[59]
2004	Physicists at the University of Manchester make graphene sheets[69]
2005	Zhang et al. demonstrate the near-field superlens—imaging objects in the tens of nanometer range with 365 nm light[61]
2007	The first left-handed material in the visible range[62]

Acknowledgments

Special thanks to Xavier Casadevall i Solvas and Drs. Sylvia Daunert and Benjamin Park.

Appendix 1A: International Technology Roadmap for Semiconductors (ITRS)

The complete 2003 ITRS and past editions of the ITRS editions are available for viewing and printing as an electronic document at http://public.itrs.net. The International Technology Roadmap for Semiconductors (ITRS) predicts the main trends in the semiconductor industry spanning across 15 years into the future. The participation of experts from Europe, Japan, Korea, and Taiwan as well as the United States ensures that the ITRS is a valid source of guidance for the semiconductor industry as it strives to extend the historical advancement of semiconductor technology and the worldwide integrated circuit (IC) market. The 2003 ITRS edition, used as the source for the tables below, extends to the year 2018. The 2003 ITRS does not predict a further acceleration in the timing of introduction of new technologies as the industry struggles through the worst recession of its history during the past couple of years. As projected, though, the half-pitch of 90 nm (hp90 nm) for DRAMs was introduced in 2004 (Intel's Prescott Pentium IV). Traditionally, the ITRS has focused on the continued scaling of CMOS (complementary metal-oxide-silicon) technology. By 2001, the horizon of the Roadmap started challenging the most optimistic projections for continued scaling of CMOS (e.g., MOSFET channel lengths below 9 nm). By that time it also became difficult for most people in the semiconductor industry to imagine how one could continue to afford the historic trends of increase in process equipment and factory costs for another 15 years! Thus, the ITRS started addressing post-CMOS devices. The Roadmap became necessarily more diverse for these devices, ranging from more familiar nonplanar CMOS devices to exotic new devices such as spintronics. Whether extensions of CMOS or radical new approaches, post-CMOS technologies must further reduce the cost per function and increase the performance of integrated circuits. Thus, new technologies may involve not only new devices but also new manufacturing paradigms.

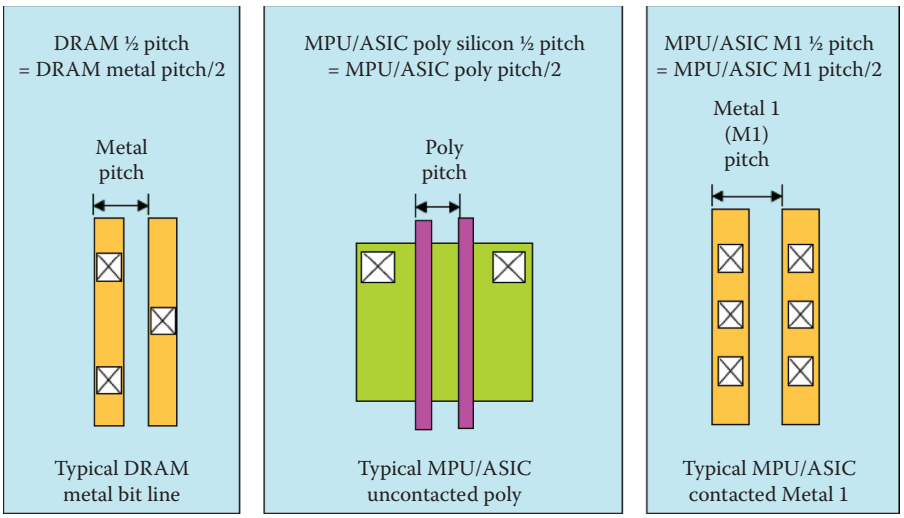
The ITRS technology nodes in the table below are defined as the minimum metal pitch used on any product, for example, either DRAM half-pitch or Metal 1 (M1) half-pitch in Logic/MPU (see also figure below the tables). In 2003, DRAMs continue to have the smallest metal half-pitch; thus, it continues to represent the technology node. Commercially used numbers for the technology generations typically differ from the ITRS technology node numbers. However, the most reliable technology standard in the semiconductor industry is provided by the above definition, which is quite clear in that the patterning and processing (e.g., etching) capabilities of the technology are represented as the pitch of the minimum metal lines. The above definition is maintained not only for the 2003 version but also as a continuation from previous ITRS editions. Therefore, the official 2003 ITRS metal hpXX node indicator has been added to differentiate the ITRS definition from commercial technology generation numbers. Interim shrink-level node trend numbers are calculated and included for convenience of monitoring the internode progress of the industry.

Near-Term Years

Year of Production	2003	2004	2005	2006	2007	2008	2009
Technology node		hp90			hp65		
DRAM half-pitch (nm)	100	90	80	70	65	57	50
MPU/ASIC MI half-pitch (nm)	120	107	95	85	75	67	60
MPU/ASIC Poly Si half-pitch (nm)	107	90	80	70	65	57	50
MPU printed gate length (nm)	65	53	45	40	35	32	28
MPU physical gate length (nm)	45	37	32	28	25	22	20

Long-Term Years

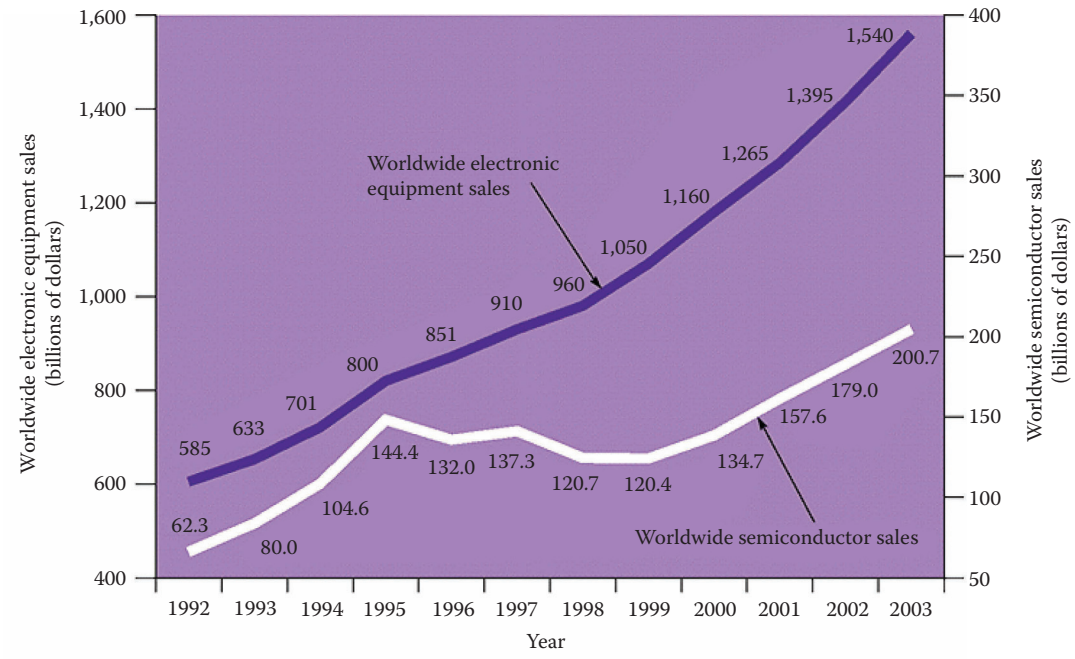
Year of Production	2010	2012	2013	2015	2016	2018
Technology node	hp45		hp32		hp22	
DRAM half-pitch (nm)	45	35	32	25	22	18
MPU/ASIC MI half-pitch (nm)	54	42	38	30	27	21
MPU/ASIC Poly Si half-pitch (nm)	45	35	32	25	22	18
MPU printed gate length (nm)	25	20	18	14	13	10
MPU physical gate length (nm)	18	14	13	10	9	7



Appendix 1B: Worldwide IC and Electronic Equipment Sales

	Amounts in US $M				Year on Year Growth in %			
	2003	2004	2005	2006	2003	2004	2005	2006
Americas	**32,330.7**	**39,514.2**	**41,734.6**	**40,089.1**	3.4	22.2	5.6	–3.9
Europe	**32,310.0**	**40,537.5**	**43,693.5**	**43,082.1**	16.3	22.5	7.8	–1.4
Japan	**38,942.2**	**47,822.9**	**51,066.8**	**50,306.9**	27.7	22.8	6.8	–1.5
Asia Pacific	**62,842.6**	**85,756.0**	**95,253.0**	**96,546.2**	22.8	36.5	11.1	1.4
Total world*	**166,425.5**	**213,630.6**	**231,748.0**	**230,024.4**	18.3	28.4	8.5	–0.7
Discrete semiconductors	**13,347.0**	**16,043.4**	**17,036.5**	**16,689.1**	8.1	20.2	6.2	–2.0
Optoelectronics	**9,544.7**	**13,100.8**	**14,851.7**	**15,281.2**	40.6	37.3	13.4	2.9
Sensors and actuators	**3,569.2**	**4,827.6**	**5,739.0**	**6,262.1**	†	35.3	18.9	9.1
Integrated circuits	**139,964.7**	**179,658.8**	**194,120.8**	**191,792.0**	16.1	28.4	8.0	–1.2
Bipolar	216.8	239.7	200.8	150.6	–4.2	10.6	–16.3	–25.0
Analog	26,793.9	33,652.2	36,971.8	36,952.4	12.0	25.6	9.9	–0.1
Micro	43,526.1	52,412.0	57,218.6	57,564.8	14.3	20.4	9.2	0.6
Logic	36,921.9	46,421.3	50,631.8	49,571.9	18.1	25.7	9.1	–2.1
Memory	32,506.0	46,933.6	49,097.9	47,552.3	20.2	44.4	4.6	–3.1
Total products*	**166,425.5**	**213,630.6**	**231,748.0**	**230,024.4**	18.3	28.4	8.5	–0.7

* All numbers are displayed as rounded to one decimal place, but totals are calculated to three decimal places precision.
† WSTS included actuators in this category from 2003. Before only sensors were reported. Therefore, a growth rate is not meaningful to show.



Percent semiconductor	10.7	12.6	14.9	18.1	15.5	15.1	12.6	11.5	11.6	12.5	12.8	13.0

Questions

1.1: Why is silicon so important to MEMS and NEMS?

1.2: Compare the pros and cons of transistors and vacuum tubes.

1.3: Why was the Si-MEMS market at one point in time expected to be much larger than the IC market?

1.4: Can you list some of the current technological and economic barriers that restrict the wider commercialization of Si-MEMS?

1.5: (a) State Moore's first law (we are talking about Moore, Intel's cofounder). (b) What is Moore's second law?

1.6: Why did surface micromachining catch on so fast with the IC industry?

1.7: Why are MEMS market forecasts so difficult to prepare? How would you go about making a better MEMS market forecast?

1.8: How does radar work? How is it useful?

1.9: What is the biggest advantage Ge has over Si IC circuits?

1.10: What is a strain gauge and what is its gauge factor?

1.11: What is the definition of nanotechnology?

1.12: List at least five commercial products that incorporate nanotechnology.

1.13: What year was the word "nanotechnology" first used?

1.14: What was Feynman's role in catalyzing the genesis of MEMS and NEMS?

1.15: Why was the honeymoon with the transistor over so quickly? What technology took over very fast?

1.16: What does ITRS stand for? What does it mean?

1.17: List a number of nanostructures that have been fabricated with bottom-up methodologies.

1.18: What is a photonic crystal?

1.19: What is a metamaterial?

1.20: What are the important differences between typical devices made in the IC industry and MEMS?

References

We took advantage of Google and Wikipedia on numerous occasions and also relied on the following references:

1. Ohl, R. S. 1946. Light-sensitive electric device, Bell Labs. US Patent 2402662.
2. Bardeen, J., and W. Brattain. 1950. Three electrode circuit element utilizing semiconductive materials, Bell Labs. US Patent 2524035.

3. Riordan, M., and L. Hoddeson. 1998. *Crystal fire: the birth of the information age.* New York: W. W. Norton & Company.
4. Reid, T. R. 2001. *The chip: how two Americans invented the microchip and launched a revolution*. New York: Random House Publishing House.
5. ITRS. 2007. International Technology Roadmap for Semiconductors. *Presentations from the 2007 ITRS Conference*, December 5, 2007, Makuhari Messe, Japan. http://www.itrs.net/Links/2007Winter/2007_Winter_Presentations/Presentations.html.
6. World Semiconductor Trade Statistics (WSTS). 2004. WSTS Semiconductor Market Forecast Spring 2004. Press release. http://www.wsts.org/plain/content/view/full/869.
7. Feynmann, R. 1959. There is plenty of room at the bottom. http://www.its.caltech.edu/~feynman/plenty.html.
8. Uhlir, A. 1956. Electrolytic shaping of germanium and silicon. *Bell Syst Tech J* 35:333–47.
9. Smith, C. S. 1954. Piezoresistance effect in germanium and silicon. *Phys Rev* 94:42–49.
10. Pfann, W. G. 1961. Improvement of semiconducting devices by elastic strain. *Solid State Electron* 3:261–67.
11. Tufte, O. N., P. W. Chapman, and D. Long. 1962. Silicon diffused-element piezoresistive diaphragms. *J Appl Phys* 33:3322–27.
12. Waggener, H. A., R. C. Kragness, and A. L. Taylor. 1967. *International Electron Devices Meeting, IEDM '67 68.* Washington, DC, IEEE.
13. Wallis, P. R., and D. I. Pomeranz. 1969. Field assisted glass-metal sealing. *J Appl Phys* 40:3946–49.
14. Shimbo, M., K. Furukawa, K. Fukuda, and K. Tanzawa. 1986. Silicon-to-silicon direct bonding method. *J Appl Phys* 60:2987–89.
15. National Semiconductor. 1974. *Transducers, Pressure, and Temperature* (catalog). Sunnyvale, CA: Author.
16. Angell, J. B., S. C. Tery, and P. W. Barth. 1983. Silicon micromechanical devices. *Scientific American* 248:44–55.
17. Greenwood, J. C. 1969. Ethylene diamine-cathechol-water mixture shows preferential etching of a p-n junction. *J Electrochem Soc* 116:1325–26.
18. Bohg, A. 1971. Ethylene diamine-pyrocatechol-water mixture shows etching anomaly in boron-doped silicon. *J Electrochem Soc* 118:401–02.
19. Texas Instruments. 1977. Texas Instruments Thermal Character Print Head. Austin, TX: Author.
20. Bassous, E., H. H. Taub, and L. Kuhn. 1977. Ink jet printing nozzle arrays etched in silicon. *Appl Phys Lett* 31:135–37.
21. O'Neil, P. 1980. A monolithic thermal converter. *Hewlett-Packard J*, 12.
22. Terry, S. C., J. H. Jerman, and J. B. Angell. 1979. A gas chromatograph air analyzer fabricated on a silicon wafer. *IEEE Trans Electron Devices* 26:1880–86.
23. Terry, S. C. 1975. *A gas chromatography system fabricated on a silicon wafer using integrated circuit technology.* PhD diss., Stanford University.
24. Petersen, K. E. 1980. Silicon torsional scanning mirror. *IBM J Res Dev* 24:631–37.
25. Petersen, K. E. 1982. Silicon as a mechanical material. *Proc IEEE* 70:420–57.
26. Howe, R. T., and R. S. Muller. 1983. Polycrystalline silicon micromechanical beams. *J Electrochem Soc* 130:1420–23.
27. Xia, Y., and G. M. Whitesides. 1998. Soft lithography. *Angew Chem Int Ed Engl* 37:551–75.
28. Nathanson, H. C., W. E. Newell, R. A. Wickstrom, and J. R. Davis. 1967. The resonant gate transistor. *IEEE Trans Electron Devices* ED-14:117–33.
29. Becker, E. W., W. Ehrfeld, D. Munchmeyer, H. Betz, A. Heuberger, S. Pongratz, W. Glashauser, H. J. Michel, and V. R. Siemens. 1982. Production of separation nozzle systems for uranium enrichment by a combination of x-ray lithography and galvanoplastics. *Naturwissenschaften* 69:520–23.
30. Fan, L. S., Y. C. Tai, and R. S. Muler. 1989. IC-processed electrostatic micro-motors. *Sensors and Actuators* A20:41–48.
31. Tang, W. C., T. C. Nguyen, and R. T. Howe. 1989. Laterally driven polysilicon resonant microstructures. *IEEE Micro Electro Mechanical Systems* 20:25–32.
32. Manz, A., N. Graber, and H. M. Widmer. 1990. Miniaturized total chemical analysis systems: A novel concept for chemical sensing. *Sens Actuators* B1:244–248.
33. Pohl, D. W., W. Denk, and M. Lanz. 1984. Optical stethoscopy: image recording with resolution l/20. *Appl Phys Lett* 44:651–53.
34. Allara, D. L., and R. G. Nuzzo. 1985. Spontaneously organized molecular assemblies; II. Quantitative infrared spectroscopic determination of equilibrium structures of solution adsorbed n-alkanoic acids on an oxidized aluminum surface. *Langmuir* 1:52–66.
35. Allara, D. L., and R. G. Nuzzo. 1985. Spontaneously organized molecular assemblies; I. Formation, dynamics and physical properties of n-alkanoic acids adsorbed from solution on an oxidized aluminum surface. *Langmuir* 1:45–52.
36. Tomalia, D. A., H. Baker, J. R. Dewald, M. Hall, G. Kallos, S. Martin, J. Roeck, J. Ryder, and P. Smith. 1985. A new class of polymers: starburst-dendritic macromolecules. *Polym J* 17:117–32.
37. Ashkin, A. 1970. Acceleration and trapping of particles by radiation pressure. *Phys Rev Lett* 24:156–59.
38. Ashkin, A., J. M. Dziedzic, J. E. Bjorkholm, and S. Chu. 1986. Observation of a single-beam gradient force optical trap for dielectric particles. *Optics Letters* 11:288–90.
39. Kroto, H. W., J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley. 1985. C60: buckminsterfullerene. *Nature* 318:162–63.
40. Tans, S. J., A. R. M. Verschueren, and C. Dekker. 1998. Room-temperature transistor based on a single carbon nanotube. *Nature* 393:49–52.
41. Ekimov, A. I., A. L. Efros, and A. A. Onushchenko. 1985. Quantum size effect in semiconductor microcrystals. *Solid State Commun* 56:921–24.
42. Efros, A. L., and A. L. Efros. 1982. Interband absorption of light in a semiconductor sphere. *Sov Phys Semicond* 16:772–74.
43. Brus, L. E. 1984. On the development of bulk optical properties in small semiconducting crystallites. *J Luminescence* 31/32: 381.
44. Brus, L. E. 1984. Electron-electron and electron-hole interactions in small semiconductor crystallites: the size dependence of the lowest excited electronic state. *J Chem Phys* 80:4403–07.
45. Murray, C. B., D. J. Norris, and M. G. Bawendi. 1993. Synthesis and characterization of nearly monodisperse CdE (E=S, Se, Te) semiconductor nanocrystallites. *J Am Chem Soc* 115:8706–15.
46. Wagner, R. S., and W. C. Ellis. 1964. Vapour-liquid-solid mechanism of single crystal growth. *Appl Phys Lett* 4:89–90.

47. Aviram, A., and M. A. Ratner. 1974. Molecular rectifiers. *Chem Phys Lett* 29:277.
48. Rayleigh, J. W. S. 1888. On the remarkable phenomenon of crystalline reflexion described by Prof. Stokes. *Phil Mag* 26:256–65.
49. Yablonovitch, E. 1987. Inhibited spontaneous emission in solid state physics and electronics. *Phys Rev Lett* 58:2059.
50. John, S. 1987. Strong localization of photons in certain disordered dielectric superlattices. *Phys Rev Lett* 58:2486.
51. Yablonovitch, E., T. J. Gmitter, and K. M. Leung. 1991. Photonic band structure: the face-centered cubic case employing nonspherical atoms. *Phys Rev Lett* 67: 2295–98.
52. Veselago, V. G. 1968. The electrodynamics of substances with simultaneously negative values of e and m. *Sov Phys Usp* 10:509–14.
53. Pendry, J. B. 1999. Photonic gap materials. *Curr Sci* 76:1311.
54. Pendry, J. B. 2000. Negative refraction makes a perfect lens. *Phys Rev Lett* 85:3966.
55. Pendry, J. B. 2004. Manipulating the near field with metamaterials. *Optics Photonics News* 15:33–37.
56. Pendry, J. B. 2007. Metamaterials and the control of electromagnetic fields. *Proceedings of the Ninth Rochester Conference on Coherence and Quantum Optics*. Washington, DC: Optical Society of America.
57. Pendry, J. B., A. J. Holden, D. J. Robbins, and W. J. Stewart. 1998. Low frequency plasmons in thin wire structures. *J Phys Cond Matter* 10:4785.
58. Ramakrishna, S. A., J. B. Pendry, M. C. K. Wiltshire, and W. J. Stewart. 2003. Imaging the near field. *J Mod Optics* 50:1419–30.
59. Wiltshire, M. C. K., J. B. Pendry, I. R. Young, D. J. Larkman, D. J. Gilderdale, and J. V. Hajnal. 2001. Microstructured magnetic material for RF flux guides in magnetic resonance imaging (MRI). *Science* 291:848–51.
60. Shelby, R. A., D. R. Smith, and S. Schultz. 2001. Experimental verification of a negative refractive index. *Science* 292:77–79.
61. Fang, N., H. Lee, and X. Zhang. 2005. Sub-diffraction-limited optical imaging with a silver superlens. *Science* 308:534–37.
62. Dolling, G., M. Wegener, C. M. Soukoulis, and S. Linden. 2007. Negative-index metamaterials at 780 nm wavelength. *Opt Lett* 32:53–55.
63. Drexel, K. E. 1987. *Engines of creation*. New York: Anchor Books.
64. Drexel, K. E. 1992. *Nanosystems, molecular machinery, manufacturing, and computation*. New York: John Wiley & Sons, Inc.
65. Atkinson, W. I. 2003. *Nanocosm: nanotechnology and the big changes coming from the inconceivable small*. New York: AMACOM.
66. Regis, E. 1995. *Nano: remaking the world atom by atom*. Boston: Little, Brown and Company.
67. Murray, C. B., C. R. Kagan, and M. G. Bawendi. 1995. Self-organization of CdSe nanocrystallites into three-dimensional quantum dot superlattices. *Science* 270:1335–38.
68. Reed, M. A., C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour. 1997. Conductance of a molecular junction. *Science* 278:252–54.
69. Geim, A. K., and K. S. Novoselov. 2007. The rise of graphene. *Nature Materials* 6:183–191.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

Crystallography

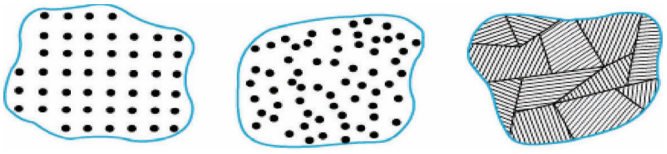
The designer Tokujin Yoshioka makes his *Venus – Natural Crystal Chair* by submerging a block of polyester fibers in the shape of a straight-backed dining chair in a vat of water and then adding a mineral to crystallize it. (Courtesy of Mr. Tokujin Yoshioka.)

Outline

Introduction

Crystallography is the science of analyzing the crystalline structure of materials. The spatial arrangement of atoms within a material plays a most important role in determining the precise properties of that material. Based on the degree of order, materials are classified as amorphous, with no recognizable long-range order; polycrystalline, with randomly ordered domains (10 Å to a few μm); and single crystalline, where the entire solid is made up of repeating units in an orderly array. This classification is illustrated in Figure 2.1. Amorphous solids (e.g., glasses and plastics) are homogeneous and isotropic because there is no long-range order or periodicity in the internal arrangement.

Many engineering materials are aggregates of small crystals of varying sizes and shapes. The size of the single-crystal grains may be as small as a few nanometers but could also be large enough to be seen by the naked eye. Regions between grains are called grain boundaries. These polycrystalline materials have properties determined by both the chemical nature of the individual crystals and their aggregate properties, such as size and shape distribution, and in the orientation relationships between them. In the case of thin polycrystalline films, material properties might deviate significantly from bulk crystalline behavior, as we discover in



(a) Crystalline (b) Amorphous (c) Polycrystalline

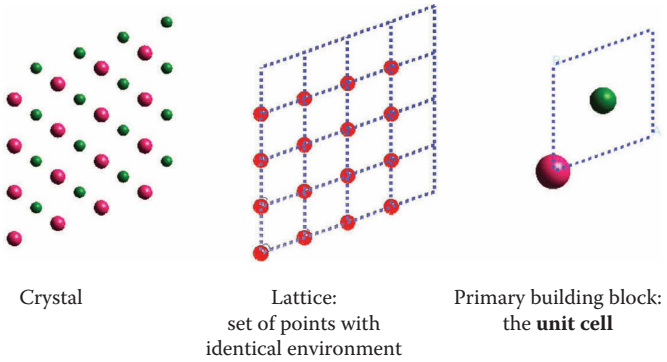
FIGURE 2.1 Classification of materials as crystalline (a), amorphous (b), and polycrystalline (c). (Drawing by Mr. Chengwu Deng.)

Volume II, Chapter 7, where we deal with thin film properties and surface micromachining. In the case of nanoparticles, deviation from expected bulk theory is even more pronounced (see Chapter 3 on quantum mechanics and the band theory of solids in the current volume). The crystal structure of a nanoparticle is not necessarily the same as that of the bulk material. Nanoparticles of ruthenium (2–3 nm in diameter), for example, have body-centered cubic (lattice point at each corner plus one at the center; see below) and face-centered cubic structures (lattice points at each corner as well as in the centers of each face; see below) not found in bulk ruthenium.

Most important, semiconductor devices are based on crystalline materials because of their reproducible and predictable electrical properties. Crystals are anisotropic—their properties vary with crystal orientation. In this chapter we explain the importance of the symmetry of point groups and space groups in determining, respectively, bulk physical properties and microscopic properties of crystalline solids, properties relied on for building miniaturized electronics, sensors, and actuators. We also launch the concept of reciprocal space (also called Fourier space, **k**-space, or momentum space), clarify the conditions for x-ray diffraction in terms of such a reciprocal space, and offer an introduction to Brillouin zones. All these elements are needed for our introduction to the band theory of solids in Chapter 3. We finish Chapter 2 with a description of crystal defects.

Bravais Lattice, Unit Cells, and the Basis

Under special conditions almost every solid can be made into a crystal (helium is the only substance that does not form a solid). Atoms organize themselves into crystals because energy can be minimized



Crystal Lattice: set of points with identical environment Primary building block: the **unit cell**

FIGURE 2.2 Any crystal lattice can be simplified to a three-dimensional array of periodically located points in space. Such a periodic array, specifying how the repeated units of a crystal are arranged, is called a *Bravais lattice*. A real crystal is made up of a basis and a lattice. (Drawing by Mr. Chengwu Deng.)

that way. Any crystal lattice can be simplified to a three-dimensional (3D) array of periodically located points in space as shown in Figure 2.2 in the case of a two-dimensional (2D) crystal. Such a periodic array, specifying how the repeated units of a crystal are arranged, is called a *Bravais lattice*. Bravais, in 1848, demonstrated that there are only 14 ways of arranging points symmetrically in space that do not lead to voids in a crystal (in 2D there are only five such lattices). All crystalline materials, including nanomaterials, assume one of the 14 Bravais lattices. A real crystal can be described in terms of a Bravais lattice, with one specific atom (or ion) or a group of atoms (a molecule), called a *basis*, attached to each lattice point (Figure 2.2). The basis superposed on the Bravais lattice renders the complete crystal structure. The 3D Bravais lattice can be mathematically defined by three noncoplanar basis vectors, $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$, which are the three independent shortest vectors connecting lattice points. These vectors form a parallelepiped called a *primitive cell*, i.e., a cell that can reproduce the entire crystal lattice by translation alone. Such a primitive cell is a minimum volume cell with a density of only one lattice point per cell—there are lattice points at each of the eight corners of the parallelepiped, but each corner point is shared among the eight cells that come together there (1/8 of a point at each corner). The lattice translational vector **r** is given by:

$$\mathbf{r} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3 \qquad (2.1)$$

where n_1, n_2, and n_3 are integers. A displacement of any lattice point by **r** will result in a new position in the lattice that has the same positional appearance as the original position. A lattice translation vector, **r**, as described in Equation 2.1, connects two points in the lattice that exhibit identical point symmetry.

Nonprimitive unit cells or simple unit cells are also called *conventional* unit cells or *crystallographic* unit cells. They are not necessarily unique and need not be the smallest cell possible. Primitive cells are chosen with the shortest possible vectors, whereas unit cells are chosen for the highest symmetry and may contain more than one lattice point per cell. The unit cell in a lattice, like a primitive cell, is representative of the entire lattice. The simplest unit cell belongs to a cubic lattice, which is further divided into simple cubic (SC), face-centered cubic (FCC), and body-centered cubic (BCC) as illustrated in Figure 2.3.

An FCC lattice has the closest atomic packing, then BCC, and then SC. For a simple cubic crystal (SC) unit cell, as shown in Figure 2.3, $\mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}_3$, and the axes angles are $\alpha = \beta = \gamma = 90°$. The dimension a ($= \mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}_3$) is known as the *lattice constant*. For SC the conventional unit cell coincides with the primitive cell. This is not true for FCC and BCC as we shall see in Figure 2.5 below.

The 14 possible Bravais lattices can be subdivided into 7 different "crystal classes" based on the choice of conventional unit cells. These 7 crystal classes are cubic, tetragonal, trigonal, hexagonal, monoclinic, orthorhombic, and triclinic. Each of these systems is characterized by a set of symmetry elements, and the more symmetry elements a crystal exhibits, obviously, the higher its symmetry. A cubic crystal has the highest possible symmetry and a triclinic crystal the lowest. The 14 Bravais lattices categorized according to the 7 crystal systems are shown in Figure 2.4.

A Wigner-Seitz cell is a primitive cell with the full symmetry of the Bravais lattice. It is an important construct for the understanding of Brillouin zones, the boundaries of which satisfy the Laue conditions for diffraction (see below). To appreciate how Wigner-Seitz cells are constructed, we illustrate some simple examples for the case of two different types of 2D lattices in Figure 2.5. Lines are drawn passing through the middle points of dotted lines connecting nearest neighbors. In 3D, the Wigner-Seitz cells

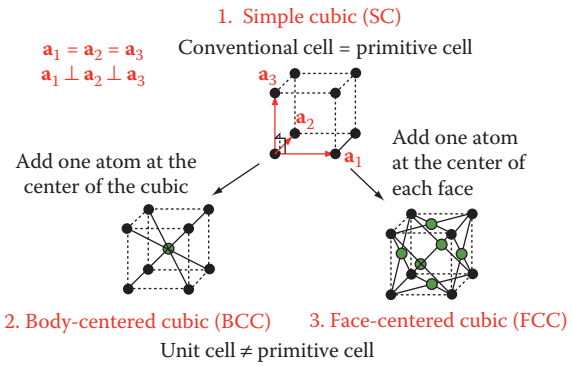


FIGURE 2.3 The simplest unit cell belongs to a cubic lattice, which is further divided into: simple cubic (SC), face-centered cubic (FCC), and body-centered cubic (BCC).

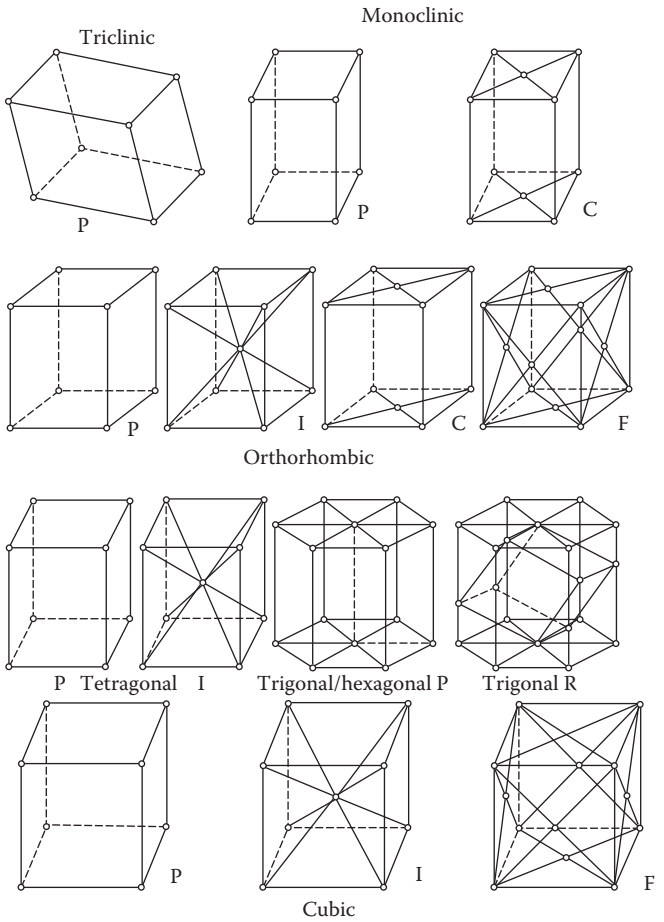


FIGURE 2.4 Conventional unit cells for the 14 Bravais lattices arranged according to the 7 crystal systems. *P* means lattice points on corners only, *C* means lattice points on corners as well as centered on faces, *F* means lattice points on corners as well as in the centers of all faces, and lattice points on corners as well as in the center of the unit cell body are indicated by *I*.

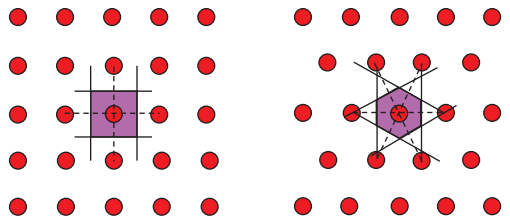


FIGURE 2.5 Wigner-Seitz primitive cells for two types of simple 2D lattices.

are polyhedra constructed about each atom by drawing planes that are the perpendicular bisectors of the lines between nearest neighbors. The Wigner-Seitz cell about a lattice point is the region of space that is closer to that point than any other lattice point. Wigner-Seitz cells for FCC and BCC Bravais lattices are shown in Figure 2.6. In the same figure we also show conventional unit and primitive unit cells for these lattices.

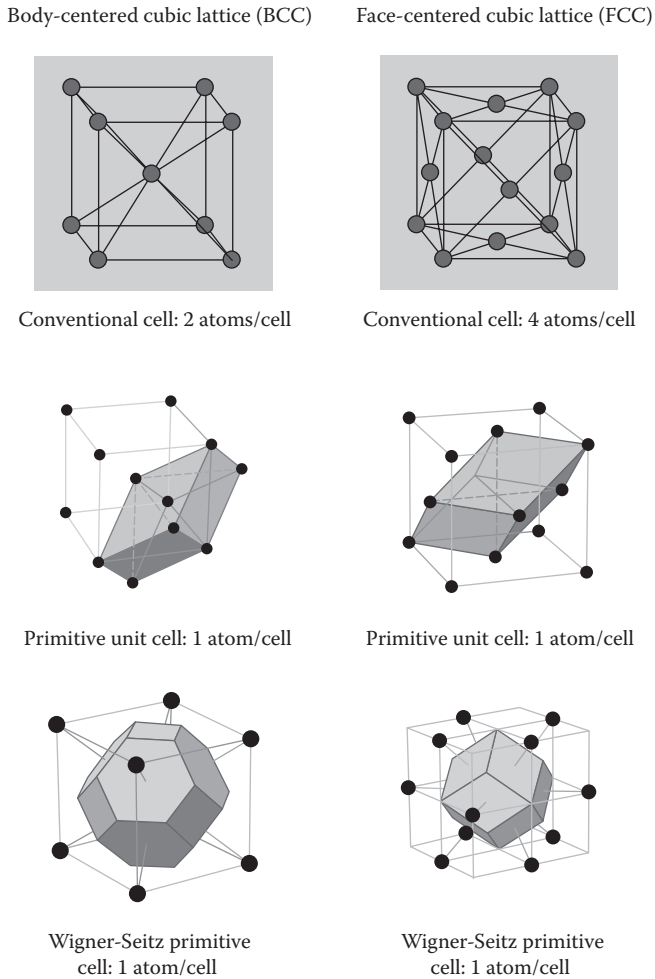


FIGURE 2.6 Conventional unit cells, primitive unit cells, and Wigner-Seitz primitive cells for BCC and FCC lattices. The BCC Wigner-Seitz unit cell is a truncated octahedron. The FCC Wigner-Seitz primitive unit cell is a rhombic dodecahedron.

Point Groups and Space Groups

A lattice translation as described by Equation 2.1 is a type of symmetry operation where a displacement of a crystal parallel to itself carries the crystal structure into itself (Figure 2.7).

Rotation and reflection or a combination of rotation and reflection—a so-called compound symmetry operation—about various points are other symmetry operations that may "carry the crystal into itself" (see Figure 2.8). The point around which the symmetry operation is carried out may be a lattice point or a special point within the elementary parallelepiped. There are five types of rotation axes possible, i.e., one- (360°), two- (180°), three- (120°), four- (90°), and sixfold (60°) rotation. One sees from Figure 2.8 why fivefold rotational symmetry does not occur in nature; it just cannot be stacked without leaving holes. This explains, for example, why we do not see ice crystals with a pentagon shape (Figure 2.9). Mirror reflection takes place about a plane through a lattice or special point. An inversion operation is an example of a compound symmetry operation and is achieved by rotation of π, followed by a reflection in a plane normal to the rotation axis; the effect is also illustrated in Figure 2.8. The collection of point symmetry elements possessed by a crystal is called a *point group* and is defined as the collection of symmetry operations which, when applied about a point, leave the lattice invariant. There are 32 crystallographic point groups in all.

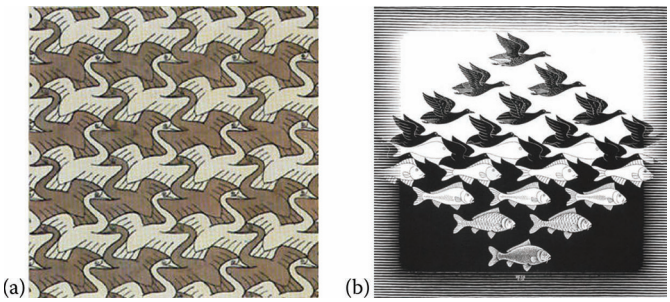


FIGURE 2.7 The drawing on the left (a) is crystal-like and can be carried into itself by a translation that is not possible in the figure on the right (b). The latter is missing a translation vector and is not crystal-like.

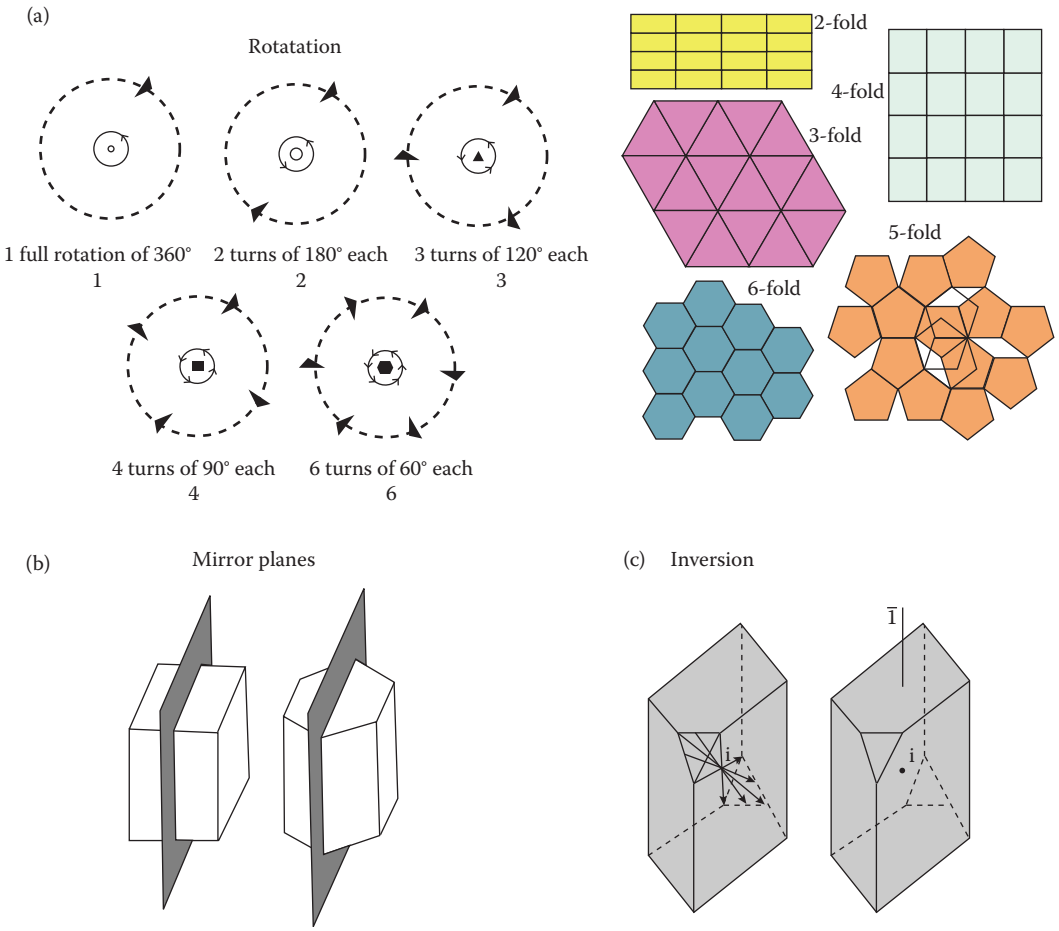


FIGURE 2.8 Point symmetry operations: (a) rotation, (b) reflection, and (c) a compound symmetry operation: inversion. The latter is made up of a rotation of π followed by reflection in a plane normal to the rotation axis. This is also called inversion through a point (i). The symbol for the inversion axis is $\bar{1}$.

The importance of the 32 point group symmetries and corresponding crystal classes is revealed by the important physical properties of crystalline solids they control, including electrical conductivity, thermal expansion, birefringence, piezoresistance, susceptibility, elastic stiffness coefficient, etc. Some properties that depend on the direction along which they are measured relative to the crystal axes are listed in Table 2.1. This orientation dependence of physical and chemical properties is called *anisotropy*. Anisotropy also explains why crystals do not grow into spheres but as polyhedra and why certain crystal directions etch faster than others. The reason for this anisotropy is the regular stacking of atoms in a crystal; as one passes along a given direction, one encounters atoms or groups of atoms at different intervals and from different angles than if one travels through the crystal from another direction. Single molecules in a liquid or a gas can also be anisotropic, but because they are free to move, liquids and gases are isotropic. In a crystal the anisotropy of atoms and groups of atoms is locked into the crystal structure. As a first example we consider a physical quantity such as current density **J** (column 4), and its cause the electrical field **E** (column 5). These quantities are linked, to a first approximation, in a linear relationship described by a tensor equation:

$$\mathbf{J} = \sigma\mathbf{E} \tag{2.2}$$

One remembers that a tensor field represents a single physical quantity that is associated with certain places in three-dimensional space and instants of time. The crystalline property, in column 1, is a tensor field of the rank listed in column 2. Also recall that a scalar (e.g., mass, temperature, charge) is a tensor of zeroth rank, and a vector (e.g., position, velocity, flow of heat) is a tensor of first rank. Other quantities such as stress inside a solid or fluid may be

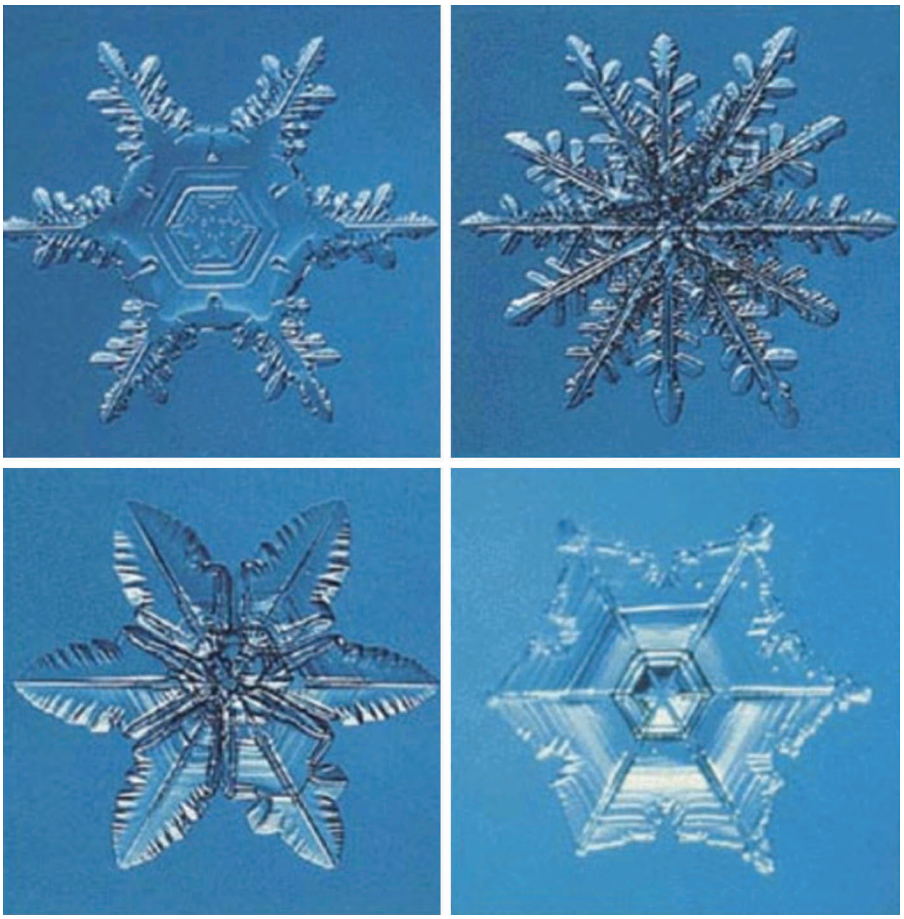


FIGURE 2.9 Ice crystals. No pentagons are found in ice crystal stacking.

characterized by tensors of order two or higher. The physical properties listed in Table 2.1 are exploited to build electronics, sensors, and actuators in MEMS and NEMS.

We elaborate a bit further here on the anisotropy of the electrical conductivity σ in a single crystal, where a causal or forcing term, the electrical field **E**, causes a current density **J**. Because both are vectors, this case is referred to as a *vector-vector* effect. In general, the current vector may not have the same direction as the electric vector. Assuming a linear relationship between electrical field (cause) and

TABLE 2.1 Linear Physical Properties of Solids

Property and Symbol	Rank of Tensor	Number of Independent Components	Dependent Physical Quantity	Causal or Forcing Term
Pyroelectric coefficient p	1	3	Electric polarization dP	Temperature change dT
Conductivity σ	2	6	Current density **J**	Field **E**
Resistivity ρ	2	6	Field **E**	Current density **J**
Susceptibility χ	2	6	Electric polarization P	Electric field **E**
Thermal expansion α	2	6	Strain ε	Temperature change dT
Piezoelectric coefficient d	3	18	Electric polarization P_s	Stress S
Elastic stiffness constants χ	4	21	Stress S	Strain ε
Elastic compliance σ	4	21	Stress ε	Stress S
Piezoresistance π	4	21	Resistivity change ρ	Stress S
Unfortunately, symbols customarily used for these properties do sometimes overlap.				

current (effect), we can describe the components of the current relative to an arbitrarily chosen Cartesian coordinate system as:

$$\begin{aligned} J_x &= \sigma_{xx}E_x + \sigma_{xy}E_y + \sigma_{xz}E_z \\ J_y &= \sigma_{yx}E_x + \sigma_{yy}E_y + \sigma_{yz}E_z \\ J_z &= \sigma_{zx}E_x + \sigma_{zy}E_y + \sigma_{zz}E_z \end{aligned} \tag{2.3}$$

The quantities σ_{ik} are components of a 3×3 "conductivity tensor." The resistivity tensor $\rho(= 1/\sigma)$ tensor, like the conductivity tensor, is a second rank tensor described by:

$$\mathbf{E} = \rho\mathbf{J} \tag{2.4}$$

Based on the basic symmetry of the equations of motion, Onsager demonstrated that the tensor is symmetric, i.e., $\sigma_{ik} = \sigma_{ki}$, so that the nine coefficients are always found to reduce to six. Taking advantage of this symmetry argument and multiplying the expressions in Equation 2.3 by E_x, E_y, and E_z, respectively, one obtains on adding:

$$\begin{aligned} J_xE_x + J_yE_y + J_zE_z &= \sigma_{xx}E^2{}_x + \sigma_{yy}E^2{}_y + \sigma_{zz}E^2{}_z \\ &+ 2\sigma_{xy}E_xE_y + 2\sigma_{yz}E_yE_z + 2\sigma_{zx}E_zE_x \end{aligned} \tag{2.5}$$

To make the mixed terms on the right side of Equation 2.5 disappear, one chooses a new coordinate system with the coordinates along the principal axes of the quadratic surface represented by this right-hand side (rhs) term, and in this new coordinate system one obtains:

$$J_x = \sigma_1E_x;\ J_y = \sigma_2E_y;\ J_z = \sigma_3E_z \tag{2.6}$$

where σ_1, σ_2, and σ_3 are the principal conductivities. The current and field vectors only have the same direction when the applied field falls along any one of the principal axes of the crystal. From Equation 2.6, no matter how low the symmetry of a crystal, it can always be characterized by three conductivities (σ_1, σ_2, and σ_3) or three specific resistivities (ρ_1, ρ_2, and ρ_3). In cubic crystals the three quantities are equal, and the specific resistivity does not vary with direction. In hexagonal, trigonal, and tetragonal crystals, two of the three principal conductivities (or resistivities) are the same. In such a case, the resistivity only depends on the angle θ between the direction in which ρ is measured and the hexagonal, trigonal, or tetragonal axis. One then finds:

$$\rho(\phi) = \rho_{per}\sin^2\phi + \rho_{par}\cos^2\phi \tag{2.7}$$

where the subscripts stand for perpendicular and parallel to the axis.

Another vector-vector effect example in Table 2.1 is the one involving thermal conductivity, where a thermal current vector is caused by a thermal gradient. Scalar-tensor effects lead to similar relations as vector-vector effects. For example, the deformation tensor of a solid resulting from a temperature change (scalar) involves three principal expansion coefficients, α_1, α_2, and α_3. The latter will again all be equal in the case of a cubic crystal, and the angular dependence of α for hexagonal, trigonal, and tetragonal crystals is given by an expression analogous to Equation 2.7.

A simple example of a scalar-vector effect from Table 2.1, illustrating the importance of crystal symmetry or lack thereof, involves pyroelectricity (see first row in Table 2.1). Pyroelectricity is the ability of a material to spontaneously polarize and produce a voltage as a result of changes in temperature. It must be a change in temperature: incident light may heat a pyroelectric crystal, thus changing its dipole moment and causing current to flow, but because pyroelectrics respond to the rate of change of temperature only, the light or heat source must be pulsed or modulated! A pyroelectric is a ferroic material, i.e., a class of smart multifunctional materials having both sensing and actuating functions. *Ferroic materials* is a simplified term to represent ferroelastic, ferromagnetic, and ferroelectric materials. In pyroelectricity, the opposite faces of certain crystals [e.g., tourmaline $(Na, Ca)(Li, Mg, Al)(Al, Fe, Mn)_6(BO_3)_3(Si_6O_{18})(OH)_4$ the "Ceylon magnet," ZnO, $BaTiO_3$, and $PbTiO_3$] become electrically charged as a result of a change in temperature. This is illustrated for tetragonal $BaTiO_3$ in Figure 2.10. In Table 2.1 we consider the linear relationship between the electric polarization **P** (a vector) and a temperature change (a scalar). Electric polarization of materials is covered from a theoretical point of view in Chapter 5. For practical applications of pyroelectricity in actuator construction refer to

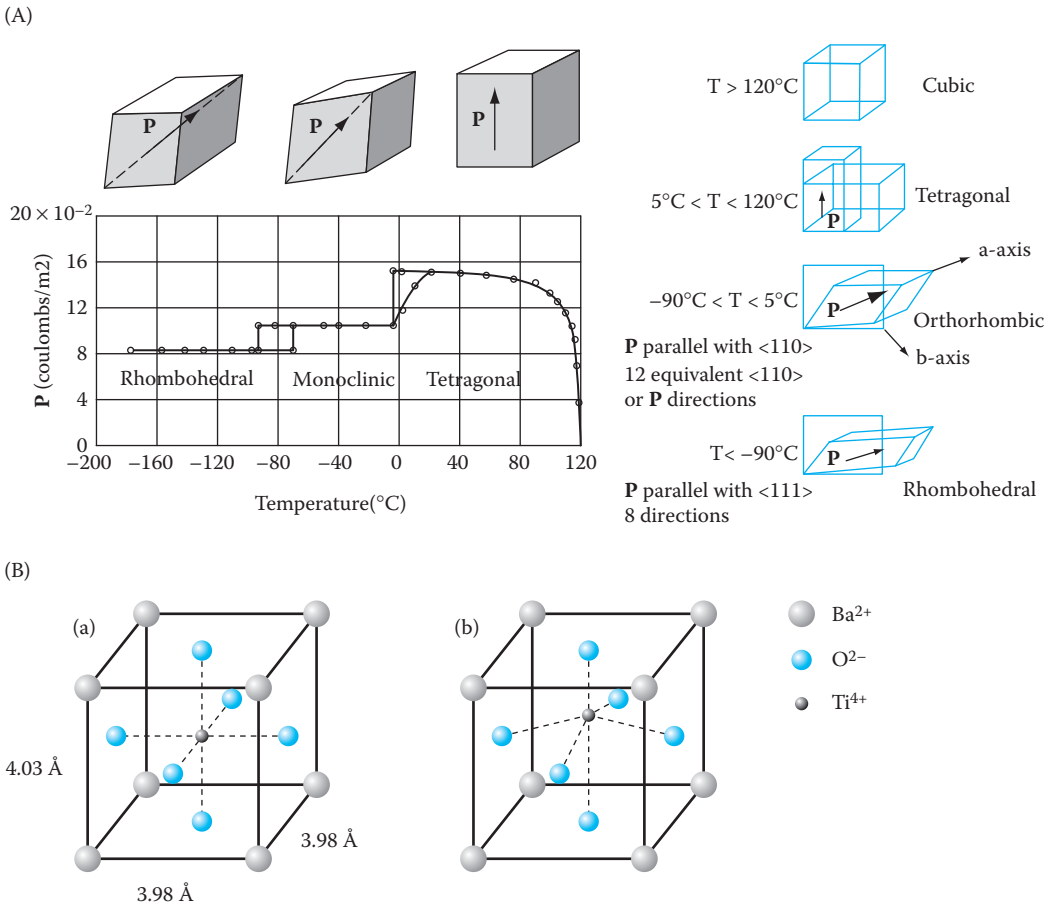


FIGURE 2.10 (A) In the pyroelectric crystal $BaTiO_3$, **P** changes with temperature only when the material is in its tetragonal state. Pyroelectricity only occurs in a crystal lacking an inversion center. This is clear from (B) (a). In cubic $BaTiO_3$ the oxygen ions are at face centers; Ba^{2+} ions are at cube corners; and Ti^{4+} is at cube center. (B) (b) In tetragonal $BaTiO_3$, the Ti^{4+} is off-center, and the unit cell has a net polarization. (Drawing by Mr. Chengwu Deng.)

Volume III, Chapter 8. The crystalline property, the pyroelectric coefficient **p**, is a tensor of rank 1 (a vector in other words). One can predict readily that pyroelectricity only occurs in crystals that lack an inversion center or a center of symmetry, i.e., in noncentrosymmetric crystals with one or more polar axes. Indeed, one could not have a crystal with one face positively charged and one negatively charged—i.e., with a polar axis—as a result of a uniform change in temperature if these crystal faces were equivalent.

More complicated cases involve vector-tensor effects and tensor-tensor effects. Piezoelectricity is an example of a vector-tensor effect; an electric field (vector) causes a mechanical deformation (tensor). Elastic deformation under the influence of a stress tensor is an example of a tensor-tensor effect. These effects require many more constants than the simple examples presented above. In discussing actuators in Volume III, Chapter 8, we will find out that an increase in symmetry introduces major simplifications in the coefficient matrices and that to describe the tensors correctly the point-group symmetry of the crystal must be known.

The combined effects of rotation or reflections from the point groups with translation from the Bravais lattice results in two additional symmetries: screw axes and glide planes. A screw axis combines rotation and translation, and a glide plane combines reflection with translation (Figure 2.11). Considering the various combinations involving the 32 point groups, screw axes, and glide planes, as well as the different Bravais lattices, a total of 230 different possible "space groups" results. In protein crystals there are only 65 space groups because all natural products are chiral so that inversion and mirror symmetry

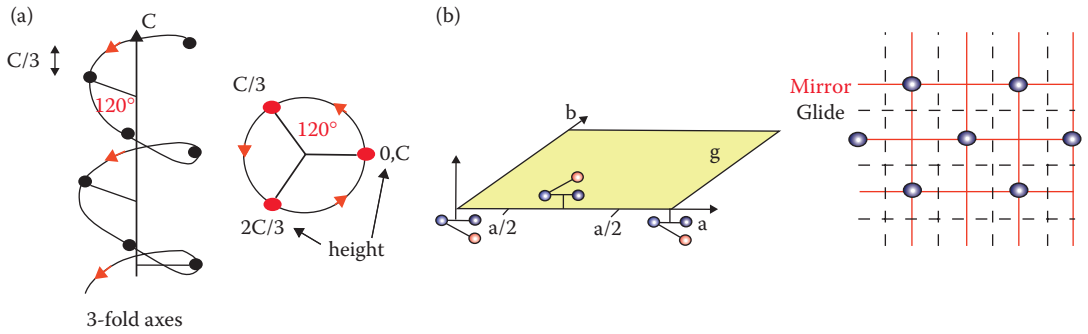


FIGURE 2.11 Example of a screw axis and a glide plane. (a) *N*-fold screw axes *C*: a combination of a rotation of 360°/*n* around *C* and a translation by an integer of *C*/*n*. (b) Glide plane: a translation parallel to the glide plane *g* by *a*/2.

operations are not allowed. A space group is a group that includes both the point symmetry elements and the translations of a crystal. These space groups are most important when studying the microscopic properties of solids. The space groups for most inorganic substances are known and can be found in tables in the Inorganic Crystal Structure Database (ICSD). These tables make it possible to calculate the exact distances and angles between different atoms in a crystal. The external shape of a crystal is referred to as the *habit*. Not all crystals have well-defined external faces. Natural faces always have low indices, i.e., their orientation can be described by Miller indices that are small integers as introduced next. The faces that we see are the lowest energy faces as the surface energy is minimized during growth.

Miller Indices

To identify a plane or a direction in a crystal, a set of integers *h*, *k*, and *l*, called the *Miller indices*, are widely used. To determine the Miller indices of a plane, one takes the intercept of that plane with the axes and expresses these intercepts as multiples of the base vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$. The reciprocal of these three integers is taken, and, to obtain whole numbers, the three reciprocals are multiplied by the smallest common denominator. The resulting set of numbers is written down as (*hkl*). By taking the reciprocal of the intercepts, infinities (∞) are avoided in plane identification. Parentheses or braces are used to specify planes.

The rules for determining the Miller indices of a direction or an orientation in a crystal are as follows: translate the orientation to the origin of the unit cell, and take the normalized coordinates of its other vertex. For example, the body diagonal in a cubic lattice as shown in the right-most panel in Figure 2.12 is 1**a**, 1**a**, and 1**a** or the [111] direction. The Miller indices for a direction are thus established using the same procedure for finding the components of a vector. Brackets or carets specify directions.

Directions [100], [010], and [001] are all crystallographically equivalent and are jointly referred to as the family, form, or group of <100> directions. A form, group, or family of faces that bear like relationships to the crystallographic axes—e.g., the planes (001), (100), (010), (00$\bar{1}$), ($\bar{1}$00), and (0$\bar{1}$0)—are all equivalent, and they are marked as {100} planes (see Figure 2.13). A summary of the typical representation for Miller indices is shown in Table 2.2. The orientation of a plane is defined by the direction

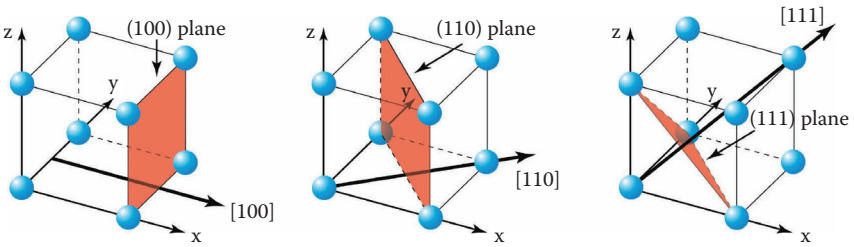


FIGURE 2.12 Miller indices for planes and directions in an SC cubic crystal. Shaded planes are from left to right (100), (110), and (111). (Drawing by Mr. Chengwu Deng.)

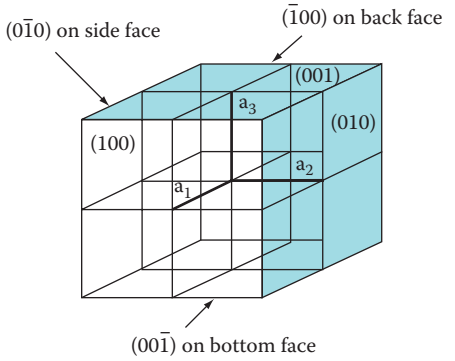


FIGURE 2.13 Miller indices for the planes of the {100} family of planes.

of a normal to the plane or the vector product ($\mathbf{A} \times \mathbf{B} = \mathbf{C}$). For a cubic crystal (such as silicon or gallium arsenide), the plane (*hkl*) is perpendicular to the direction [*hkl*]. In other words, the indices of a plane are the same numbers used to specify the normal to the plane. Using a simple cubic lattice as an example, you can check that crystal planes with the smallest Miller indices, such as {100}, {110}, {111}, have the largest density of atoms. Usually crystals are cleaved along these planes and are grown in directions perpendicular to them.

When one comes across more complicated planes than the ones considered above, the mathematical vector algebra approach to calculate the Miller indices becomes useful. For examples, consider the plane in Figure 2.14, defined by three points P1, P2, and P3, where P1: (400), P2: (020), and P3: (003).

Step 1. Define the following vectors:

$$\mathbf{r} = x\mathbf{a}_1 + y\mathbf{a}_2 + z\mathbf{a}_3$$

$$\mathbf{r}_1 = 4\mathbf{a}_1 + 0\mathbf{a}_2 + 0\mathbf{a}_3$$

$$\mathbf{r}_2 = 0\mathbf{a}_1 + 2\mathbf{a}_2 + 0\mathbf{a}_3 \qquad (2.8)$$

$$\mathbf{r}_3 = 0\mathbf{a}_1 + 0\mathbf{a}_2 + 3\mathbf{a}_3$$

TABLE 2.2 Miller Indices Symbols

Notation	Interpretation
(*hkl*)	Crystal plane
{*hkl*}	Equivalent planes
[*hkl*]	Crystal direction
<*hkl*>	Equivalent directions

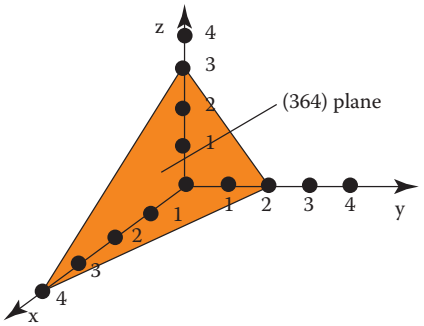


FIGURE 2.14 The (364) plane in a SC cubic lattice.

and find the differences:

$$\mathbf{r} - \mathbf{r}_1 = (x - 4)\,\mathbf{a}_1 + (y - 0)\,\mathbf{a}_2 + (z - 0)\,\mathbf{a}_3$$

$$\mathbf{r}_2 - \mathbf{r}_1 = (0 - 4)\,\mathbf{a}_1 + (2 - 0)\,\mathbf{a}_2 + (0 - 0)\,\mathbf{a}_3 \qquad (2.9)$$

$$\mathbf{r}_3 - \mathbf{r}_1 = (0 - 4)\,\mathbf{a}_1 + (0 - 0)\,\mathbf{a}_2 + (3 - 0)\,\mathbf{a}_3$$

Step 2. Calculate the scalar triple product of these three vectors [$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C})$], which in this case is a plane and its volume is zero because the vectors are coplanar [$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = 0$]:

$$(\mathbf{r} - \mathbf{r}_1) \cdot [(\mathbf{r}_2 - \mathbf{r}_1) \times (\mathbf{r}_3 - \mathbf{r}_1)] = 0 \qquad (2.10)$$

For vectors **A**, **B**, **C** with coordinates (A_1, A_2, A_3), (B_1, B_2, B_3), and (C_1, C_2, C_3), the requirement in Equation 2.10 is equivalent to:

$$\mathbf{A} \cdot \mathbf{B} \times \mathbf{C} = \begin{vmatrix} A_1 & A_2 & A_3 \\ B_1 & B_2 & B_3 \\ C_1 & C_2 & C_3 \end{vmatrix} = A_1(B_2C_3 - B_3C_2)$$

$$+ A_2(B_3C_1 - B_1C_3) + A_3(B_1C_2 - B_2C_1) \qquad (2.11)$$

For our example this leads to:

$$\begin{vmatrix} x - 4 & y & z \\ -4 & 2 & 0 \\ -4 & 0 & 3 \end{vmatrix} = (x - 4)6 + 12y + 8z = 0$$

or $3x + 6y + 4z = 12$

Once we have the equation for the plane, we easily find the Miller indices.

Step 3. Determine the Miller indices.

1. The intercepts with the axes: $x = 4$ (for $y = z = 0$), $y = 2$ (for $x = z = 0$), and $z = 3$ (for $x = y = 0$)
2. The reciprocals 1/4, 1/2, and 1/3, or
3. The Miller indices for the plane are (364)

Adjacent planes (*hkl*) in a simple cubic crystal are spaced a distance d_{hkl} from each other, with d_{hkl} given by:

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \tag{2.12}$$

where *a* is the lattice constant. Equation 2.12 provides the magnitude of d_{hkl} and follows from simple analytic geometry. To generalize this expression, notice that for a plane (*hkl*) with $hx + ky + lz = a$, the distance from any point (x_1, y_1, z_1) to this plane is:

$$d_{hkl} = \left| \frac{hx_1 + ky_1 + lz_1 - a}{\left(h^2 + k^2 + l^2\right)^{\frac{1}{2}}} \right| \tag{2.13}$$

Hence when that point is at origin (0, 0, 0) we find Equation 2.12 back.

- Example 2.1: With a = 5 Å, we find d = a = 5 Å for (100) planes and d = a/$\sqrt{2}$ = 3.535 Å for (110) planes.

Because a, b = |a||b| cos θ the angle between plane (h_1, k_1, l_1) and plane (h_2, k_2, l_2) can be calculated as:

$$\cos\theta = \frac{(h_1h_2 + k_1k_2 + l_1l_2)}{\sqrt{h_1^2 + k_1^2 + l_1^2}\sqrt{h_2^2 + k_2^2 + l_2^2}} \tag{2.14}$$

- Example 2.2: The angle between *a* (100) and *a* (111) plane is cos θ = 1/$\sqrt{1}\sqrt{3}$ = 0.58 or θ = 54.74°.

X-Ray Analysis

Introduction

X-ray analysis reveals the symmetries of crystals (lattice type), distances between atomic planes (lattice parameter), the positions of atoms in crystals, the types of atoms from the intensities of diffracted x-rays, and the degree of crystallinity (ordering). To perform x-ray crystallography, it is necessary to grow crystals with edges of around 0.1–0.3 mm. This is usually not a problem for inorganic materials, but in the case of organic materials, such as proteins (for example, see the crystal structure of the GFP protein in Figure 7.108) and nucleic acids (see x-ray diffraction image in Figure 2.17), it often is a challenge: imagine trying to crystallize a molecule with 10,000 atoms! A crystallographer must combine ingenuity and patience to trick these molecules into crystallizing. In this section we learn about diffraction and the all-important Bragg and Laue x-ray diffraction laws and how the latter are used to deduce the 3D structure of molecules.

Fourier Transforms

Diffraction forms the basis for x-ray crystallography. The first step toward interpreting diffraction patterns was a mathematical trick discovered by the French mathematician Joseph Fourier, who in 1807 introduced Fourier transforms for solving heat conduction problems. The result of a Fourier transform is that periodic functions in the time domain, e.g., light waves, can be completely characterized by information in the frequency domain, i.e., by frequencies and amplitudes of sine, cosine functions. Fourier analysis provides us with the tools to express most functions as a superposition of sine and cosine waves of varying frequency. For periodic signals a discrete sum of sines/cosines of different frequencies is multiplied by a different weighting coefficient in a so-called Fourier series (FS). For nonperiodic functions, one needs a continuous set of frequencies so the integral of sines/cosines is multiplied by a weighting function in a so-called Fourier transform (FT) (Figure 2.15). One important property of Fourier transforms is that they can be inverted. If you apply a Fourier transform to some function, you can take the result and run it through an inverse Fourier transform to get the original function back. The inverse Fourier transform is essentially just another Fourier transform. Fourier and inverse Fourier transforms, which take the signal back and forth between time and frequency domains, are:

$$\text{From time to frequency: } X(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t)\, e^{-2\pi i f t} dt \tag{2.15}$$

$$\text{From frequency to time: } x(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} X(f)\, e^{2\pi i f t} df \tag{2.16}$$

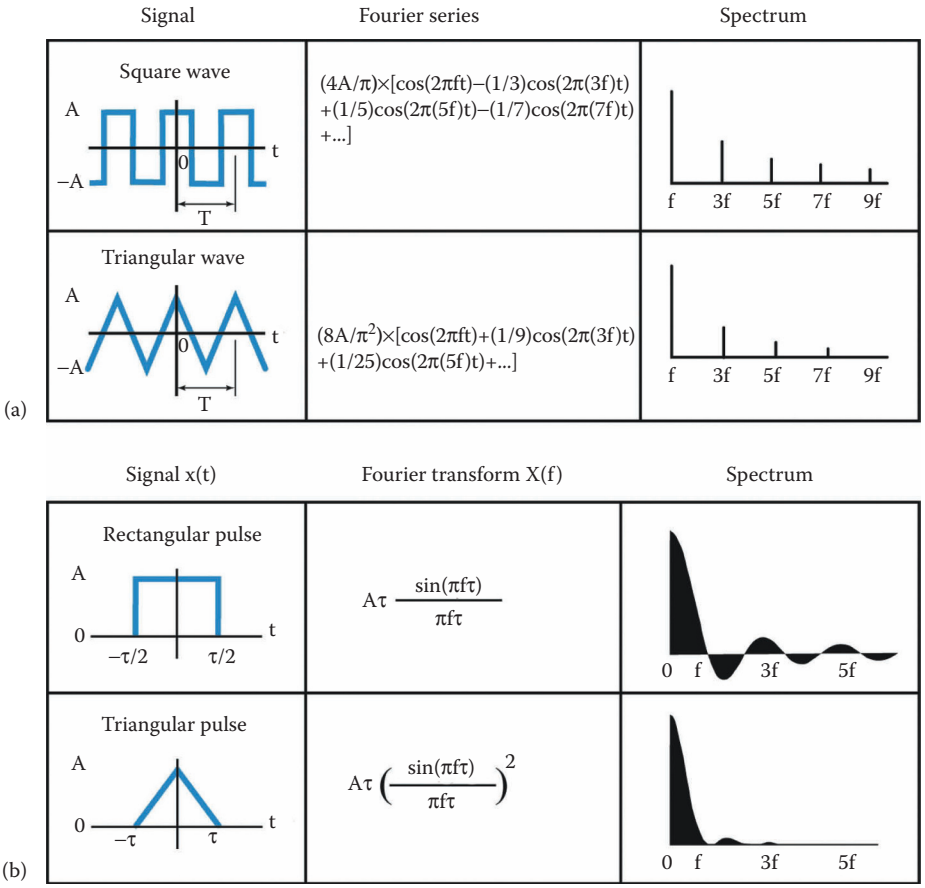


FIGURE 2.15 (a) For periodic signals a discrete sum of sines/cosines of different frequencies is multiplied by a different weighing coefficient in a so-called Fourier series (FS). (b) For nonperiodic functions, one needs a continuous set of frequencies so the integral of sines/cosines is multiplied by a weighting function in a so-called Fourier transform (FT).

In Chapter 5 we will see that the formation of an image, according to Abbe's theory, is a two-stage, double-diffraction process: an image is the diffraction pattern of the diffraction pattern of an object. In x-ray diffraction there is no lens to focus the x-rays, so we have to use a computer to reassemble the image: the x-ray diffraction patterns from a crystal are related to the object diffracting the waves through a Fourier transform.

Fourier transforms are actually even more general than revealed here: a FT allows for a description given in one particular "space" to be transformed to a description in the reciprocal of that space, and time-frequency transformation is just one example. It is interesting to note that strong criticism by peers blocked publication of Fourier's work until 1822 (*Theorie Analytique De La Chaleur*). Today, Fourier analysis is used in GSM (global system for mobile communications)/cellular phones, most DSP (digital signal processing)-based applications, music, audio, accelerator control, image processing, x-ray spectrometry, chemical analysis (FT spectrometry), radar design, PET scanners, CAT scans and MRI, speech analysis (e.g., voice-activated "devices," biometry), and even stock market analysis.

X-Ray Diffraction

Introduction

Wilhelm Conrad Roentgen discovered x-rays in 1895 and received the Nobel Prize in Physics in 1901 for his discovery. X-rays are scattered by the electrons in atoms because electromagnetic radiation (including x-rays) interacts with matter through its fluctuating electric field, which accelerates charged particles. You can think of electrons oscillating in position and, through their accelerations, re-emitting electromagnetic radiation. The scattered radiation

interferes both constructively and destructively, producing a diffraction pattern that can be recorded on a photographic plate. For x-rays, electrons, and neutrons incident on a single crystal, diffraction occurs because of interference between waves scattered elastically from the atoms in the crystal. Intensity of scattered radiation is proportional to the square of the charge/mass ratio, and the proton is about 2000 times as massive as the electron. Because electrons have a much higher charge-to-mass ratio than atomic nuclei or even protons, they are much more efficient in this process. With x-rays the interaction is with the electron mantle of the atoms. In the case of electron beams, say in an electron microscope, scattering is from both the electron mantle and the atom nuclei, and neutrons interact with the nucleus only.

The final result of a crystallographic experiment is not a picture of the atoms, but a map of the distribution of electrons in the molecule, i.e., an electron density map (Figure 2.16). Because the electrons are mostly tightly localized around the nuclei, the electron density map gives us a pretty good picture of the molecule. As we do not have a lens we do not get the electron density map directly; the x-ray diffraction patterns from a crystal are related to the object diffracting the waves through a Fourier transform.

If one thinks of electron density as a mathematical function:

$$\rho(x, y, z) \tag{2.17}$$

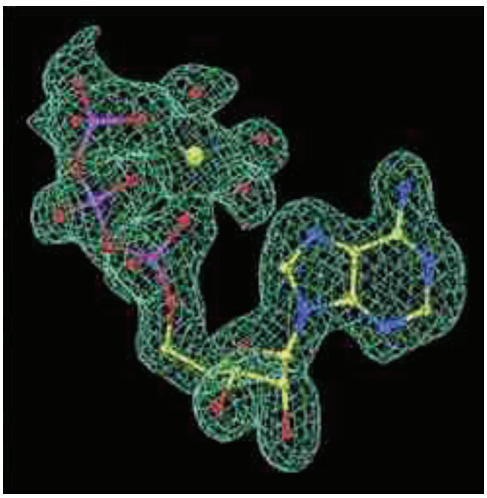


FIGURE 2.16 Electron density map of adenosine triphosphate (ATP).

with x, y, z indices for real space, then the diffraction pattern is the Fourier transform of that electron density function and given as:

$$F(hkl) = T[\rho(x, y, z)] \tag{2.18}$$

where $F(hkl)$ is the structure factor (a scattered wave, therefore a complex number with amplitude and phase) with hkl indices in reciprocal space, and T is the forward Fourier transform of $\rho(x, y, z)$. As we saw in Equation 2.16, the reverse relationship holds also, namely:

$$\rho(x, y, z) = T^{-1}\{T[\rho(x, y, z)]\} \tag{2.19}$$

where T^{-1} is the inverse Fourier transform. This expression tells us that the inverse Fourier transform of the Fourier transform of an object is the original object. The latter is a rewording of Abbe's treatment of image generation as presented in Chapter 5. The intensity in an x-ray diffraction photograph is the square of the amplitude of the diffracted waves, $|F(X)|^2$, or the recorded diffraction pattern of an object is the square of the Fourier transform of that object. In a diffraction pattern, each point arises from the interference of rays scattered from all irradiated portions of the object. To determine the image, one must measure or calculate the structure factor $F(X)$ at many or all points of the diffraction pattern. Each $F(X)$ is described by an amplitude and a phase, but in recording the intensity of the diffracted x-rays, the phase information is lost. This is referred to as the *phase* problem. X-ray phases could be obtained directly if it were possible to rediffract (focus) the scattered rays with an x-ray lens to form an image; unfortunately, an x-ray lens does not exist. With x-rays we can thus detect diffraction from molecules, but we have to use a computer to reassemble an image as shown in Figure 2.16. The process is summarized in Figure 2.17. We will now learn how the x-ray diffraction pattern, $F(X)$, comes about and how the expected intensities in the diffraction pattern are calculated to solve Equation 2.19, i.e., how to reconstitute the image that led to the measured x-ray diffraction pattern.

Bragg's Law

In 1913 W.H. and W.L. Bragg (a father and son team) proposed that the condition for constructive

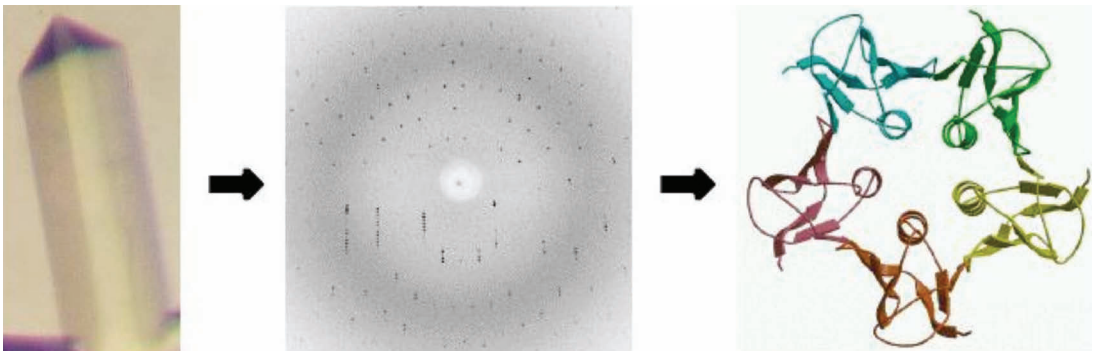


FIGURE 2.17 With x-rays, we can detect diffraction from molecules, but we have to use a computer to reassemble the electron density/molecular structure image.

specular reflection of x-rays from a set of crystal planes separated by a distance d_{hkl} could be represented as:

$$2d_{hkl} \sin \theta = n\lambda \tag{2.20}$$

This expression basically tells us that constructive interference of waves reflected by successive crystal planes occurs whenever the path difference ($2d_{hkl}\sin\theta$) is an integral multiple (n) of the wavelength λ. Also, for each (hkl) family of planes, x-rays will only diffract at one angle θ. The integer n is known as the order of the corresponding reflection. Because Bragg reflection can only occur for $\lambda \leq 2d$, one needs x-rays with wavelengths in the Ångstrom range to resolve crystal planes. The Bragg equation is easily derived from an inspection of Figure 2.18. Bragg's law is a result of the periodicity of the lattice with the atoms in the crystal basis controlling the relative intensity of the various orders (n) of diffraction from a set of parallel (hkl) planes. This basic equation is the starting point for understanding crystal diffraction of x-rays, electrons, neutrons, and any other particles that have a de Broglie wavelength

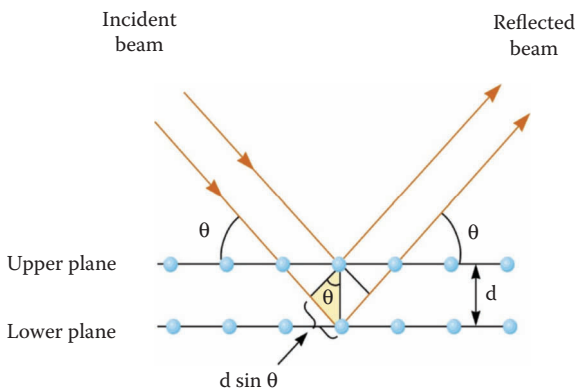


FIGURE 2.18 Schematic used to derive the Bragg equation.

(Chapter 3) less than the interatomic spacing. Although the reflection from each plane is specular, only for certain values of θ will the reflections from all planes add up in phase to give a strong reflected beam. Each plane reflects only 10^{-3} to 10^{-5} of the incident radiation, i.e., it is not a perfect reflector. Hence, 10^3 to 10^5 planes contribute to the formation of the Bragg-reflected beam in a perfect crystal.

The composition of the basis determines the relative intensity of the various orders of diffraction.

Laue Equations

In 1912 von Laue predicted that diffraction patterns of x-rays on crystals would be entirely analogous to the diffraction of light by an optical grating.* In the von Laue approach there is no ad hoc assumption of specular reflection, as in the case of Bragg's law. Instead, this more general approach considers a crystal as composed of sets of ions or atoms at the sites of a Bravais lattice that reradiate the incoming x-rays. For both optical gratings and crystals only the repeat distances of the periodic structure and the wavelength of the radiation determine the diffraction angles.

Let us inquire first into the interference conditions for waves originating from different but identical atoms in a single row—the one-dimensional diffraction case. The scattering atoms in a line form secondary, coherent x-ray sources (scattering from two atoms is shown in Figure 2.19).

* We will encounter the diffraction of light by an optical grating again in Volume II, Chapter 1 on photolithography, where we discuss patterning of a photoresist with UV light, using a mask with a grating structure on it.

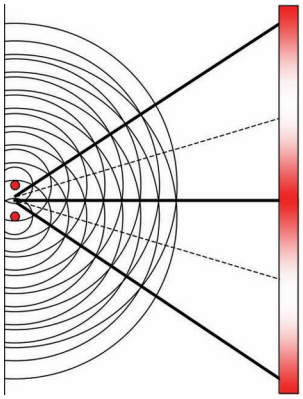


FIGURE 2.19 Two scattering atoms act as coherent secondary sources.

Constructive interference will occur in a direction such that contributions from each lattice point differ in phase by 2π. This is illustrated for the scattering of an incident x-ray beam by a row of identical atoms with lattice spacing $\mathbf{a}_1$ in Figure 2.20. The direction of the incident beam is indicated by wave vector $\mathbf{k}_0$ or the angle α_0, and the scattered beam is specified by the direction of $\mathbf{k}$ or the angle α. Because we assume elastic scattering, the two wave vectors $\mathbf{k}_0$ and $\mathbf{k}$ have the same magnitude, i.e., $2\pi/\lambda$ but with differing direction. A plane wave* $e^{ik.r}$ is constant in a plane perpendicular to $\mathbf{k}$ and is periodic parallel to it, with a wavelength $\lambda = 2\pi/\mathbf{k}$ (see Appendix 2A). The path difference $A_1B - A_2C$ in Figure 2.20 must equal $e\lambda$ with e = 0, 1, 2, 3, …. For a fixed incident x-ray with wavelength λ and direction $\mathbf{k}$, and an integer value of e, there is only one possible scattering angle α defining a cone of rays drawn about a line through the lattice points (see Figure 2.20). Because crystals are periodic in three directions, the Laue equations in 3D are then:

$$\begin{aligned} \mathbf{a}_1(\cos\alpha - \cos\alpha_0) &= e\lambda \\ \mathbf{a}_2(\cos\beta - \cos\beta_0) &= f\lambda \\ \mathbf{a}_3(\cos\gamma - \cos\gamma_0) &= g\lambda \end{aligned} \qquad (2.21)$$

For constructive interference from a three-dimensional lattice to occur, the three equations above must all be satisfied simultaneously, i.e., six angles α, β, γ, a, α_0, β_0, and γ_0; three lattice lengths $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$; and three integers (e, f, and g) are fixed. Multiplying both sides of Equation 2.21 with $2\pi/\lambda$ and rewriting the expression in vector notation we obtain:

$$\begin{aligned} \mathbf{a}_1 \cdot (\mathbf{k} - \mathbf{k}_0) &= 2\pi e \\ \mathbf{a}_2 \cdot (\mathbf{k} - \mathbf{k}_0) &= 2\pi f \\ \mathbf{a}_3 \cdot (\mathbf{k} - \mathbf{k}_0) &= 2\pi g \end{aligned} \qquad (2.22)$$

with $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ being the primitive vectors of the crystal lattice. When two of the conditions in Equation 2.22 are met, one entire plane array scatters in phase. This is depicted in Figure 2.21, where two cones of allowed diffracted rays are depicted. The two conditions are met simultaneously only in two directions along which the cones intersect. To satisfy all three Laue equations simultaneously, the diffracted beam can only have one allowed direction because three cones can mutually intersect along only one line.

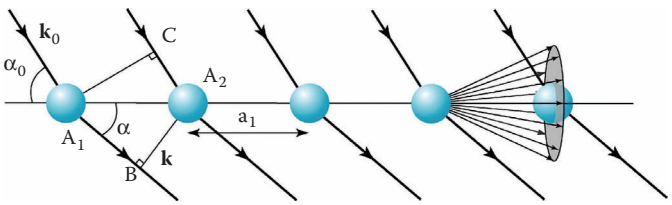


FIGURE 2.20 Scattering of an incident x-ray beam (incident direction is $\mathbf{k}_0$) by a row of identical atoms with lattice spacing $\mathbf{a}_1$. The scattered beam is specified by the direction $\mathbf{k}$. The path difference $A_1B - A_2C$ must equal $e\lambda$, with e = 0,1,2,3,.... (Drawing by Mr. Chengwu Deng.)

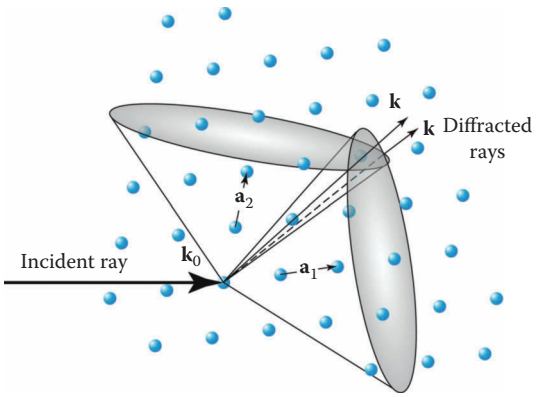


FIGURE 2.21 Each Laue condition produces a cone of allowed rays. In a plane array the entire plane scatters in phase in two directions. These two directions are along the intersection of the two cones. (Drawing by Mr. Chengwu Deng.)

* The complex exponential representation for periodic functions is convenient for adding waves, taking derivatives of wave functions, and so on. It is equivalent to a linear combination of a sine and cosine function, because $e^{i\theta} = \cos\theta + i\sin\theta$ (Euler) (see also Appendix 2A).

FIGURE 2.22 Max von Laue (1897–1960).

If we further define a vector $\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$, Equation 2.22 simplifies to:

$$\begin{aligned} \mathbf{a}_1 \cdot \Delta\mathbf{k} &= 2\pi e \\ \mathbf{a}_2 \cdot \Delta\mathbf{k} &= 2\pi f \\ \mathbf{a}_3 \cdot \Delta\mathbf{k} &= 2\pi g \end{aligned} \qquad (2.23)$$

Dealing with 12 variables for each reflection simultaneously [six angles (α, β, γ, a, α_0, β_0, and γ_0), three lattice lengths ($\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$), and three integers (e, f, and g)] is a handful; this is the main reason why the Laue equations are rarely referred to directly, and a simpler representation is used instead. The reflecting conditions can indeed be described more simply by the Bragg equation. Historically, von Laue (Figure 2.22) developed his equations first; it was one year after his work that the father and son team William Henry and William Lawrence Bragg (Figure 2.23) introduced

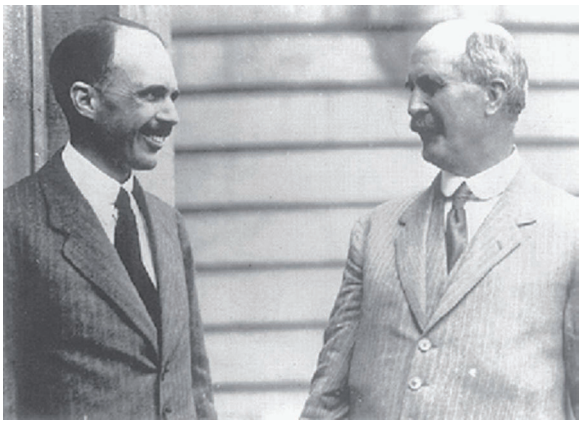


FIGURE 2.23 Father and son Bragg: Sir William Henry and William Lawrence Bragg.

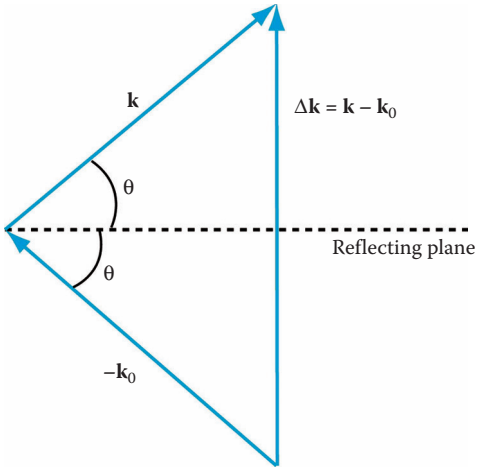


FIGURE 2.24 In case of mirror-like Bragg reflection, the vector **Δk**, the summation of the unit vectors representing incoming ($\mathbf{k}_0$) and reflected rays (**k**), is normal to the plane that intersects the 2θ angle between them. (Drawing by Mr. Chengwu Deng.)

the simpler Bragg's law. Max von Laue and the Braggs received the Nobel Prize in Physics in 1914 and 1915, respectively.

Further below we will learn that constructive interference of diffracted x-rays will occur provided that the change in wave vector, $\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$, is a vector of the reciprocal lattice.

Bragg's law is equivalent to the Laue equations in one dimension as can be appreciated from an inspection of Figures 2.24 and 2.25, where we use a two-dimensional crystal for simplicity. Suppose that vector $\Delta\mathbf{k}$ in Figure 2.24 satisfies the Laue condition; because incident and scattered waves have the same magnitude (elastic scattering), it follows that incoming ($\mathbf{k}_0$) and reflected rays (**k**) make the same angle θ with the plane perpendicular to $\Delta\mathbf{k}$.

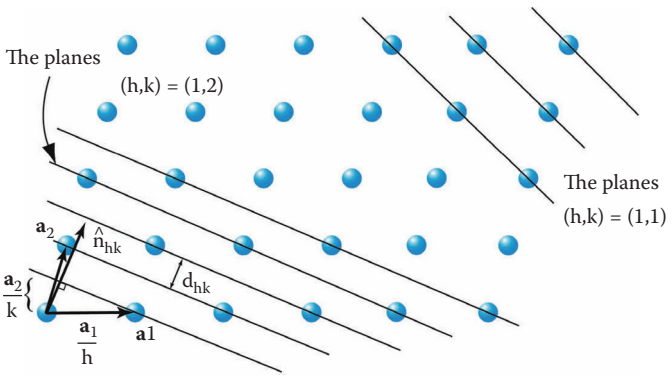


FIGURE 2.25 Connecting Bragg's law with Laue equations and Miller indices. (Drawing by Mr. Chengwu Deng.)

The magnitude of vector $\Delta\mathbf{k}$, from Figure 2.24, is then given as:

$$|\Delta\mathbf{k}| = 2\mathbf{k}\sin\theta \qquad (2.24)$$

We now derive the relation between the reflecting planes to which $\Delta\mathbf{k}$ is normal and the lattice planes with a spacing d_{hkl} (see Figure 2.25 and Bragg's law in Equation 2.20). The normal unit vector $\hat{\mathbf{n}}_{hk}$ and the interplanar spacing d_{hk} in Figure 2.25 characterize the crystal planes (*hk*). From Equation 2.23 we deduce that the direction cosines of $\Delta\mathbf{k}$, with respect to the crystallographic axes, are proportional to e/a_1, f/a_2, and g/a_3 or:

$$e/a_1{:}f/a_2{:}g/a_3 \qquad (2.25)$$

From the definition of the Miller indices, an (*hkl*) plane intersects the crystallographic axes at the points a_1/h, a_2/k, and a_3/l, and the unit vector $\hat{\mathbf{n}}_{hkl}$, normal to the (*hkl*) plane, has direction cosines proportional to:

$$h/a_1,\ k/a_2,\ \text{and}\ l/a_3 \qquad (2.26)$$

Comparing Equations 2.25 and 2.26 we see that $\Delta\mathbf{k}$ and the unit normal vector $\hat{\mathbf{n}}_{hkl}$ have the same directions; all that is required is that e = nh, f = nk, and g = nl, where *n* is a constant. The factor *n* is the largest common factor of the integers *e*, *f*, and *g* and is itself an integer. From the above, Laue's equations can also be interpreted as reflection from the *h,k,l* planes. From Figure 2.25 it can be seen that the spacing between the (*hk*) planes, and by extension between (*hkl*) planes, is given as*:

$$d_{hkl} = \frac{\hat{\mathbf{n}}_{hkl}\cdot\mathbf{a}_1}{h} = \frac{\hat{\mathbf{n}}_{hkl}\cdot\mathbf{a}_2}{k} = \frac{\hat{\mathbf{n}}_{hkl}\cdot\mathbf{a}_3}{l} \qquad (2.27)$$

Because $\Delta\mathbf{k}$ is in the direction of the normal $\hat{\mathbf{n}}_{hkl}$ and comes with a magnitude given by Equation 2.24, we obtain Bragg's law from the Laue's equations as:

$$\mathbf{a}_1\cdot\Delta\mathbf{k} = \mathbf{a}_1\cdot\hat{\mathbf{n}}_{hkl}\,2k\sin\theta = 2\pi e \qquad (2.28)$$

or:

$$hd_{hkl}\frac{4\pi}{\lambda}\sin\theta = 2\pi e \qquad (2.29)$$

* Notice that in the case of a cubic crystal, Equation 2.27 can be simplified to give the distance between planes as in Equation 2.12.

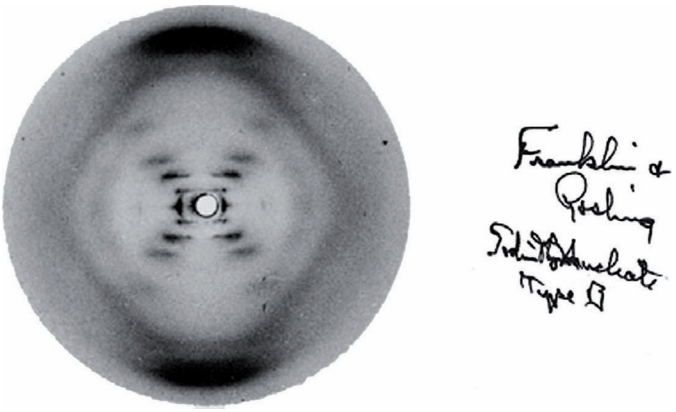


FIGURE 2.26 Sodium deoxyribose nucleate from calf thymus. (Structure B, Photo 51, taken by Rosalind E. Franklin and R.G. Gosling.) Linus Pauling's annotations are to the right of the photo (May 2, 1952).

and with e = nh:

$$2d_{hkl}\sin\theta = n\lambda \qquad (2.30)$$

In the Bragg equation we treat x-ray diffraction from a crystal as a reflection from reciprocal lattice planes rather than scattering from atoms. This construction has fewer variables than the Laue equations because reflections are wholly represented in two dimensions only.

In Figure 2.26 we reproduce what is possibly the most famous x-ray diffraction photograph. It was this photograph—photo 51 of DNA taken by Rosalind Franklin and R.G. Gosling—that convinced Watson and Crick that the DNA molecule was helical. The discovery of the double helix followed soon after, as well as an enduring controversy that Franklin probably did not get the credit she deserved in the elucidation of the DNA structure.

X-Ray Intensity and Structure Factor F(hkl)

So far we have considered only the condition for diffraction from simple lattices for which only corner points of the unit cell are occupied. The intensity of a beam diffracted from an actual crystal will depend on the grouping of atoms in the unit cell and on the scattering power of these atoms. In this section we discuss the intensity and phase of the diffracted rays and the structure factor, *F*(*hkl*). If we treat the incident x-ray waves as plane waves and the atoms as ideal point scatterers, the scattered waves are spherical waves (Figure 2.19) close to the source, i.e., near-field or Fresnel diffraction patterns form at finite

distances from the crystal and Fraunhofer diffraction patterns at infinity (far-field). In a typical x-ray diffraction experiment, a Fraunhofer diffraction pattern is registered at about 50 to 150 mm behind the specimen. This distance is relatively large compared to the size of the diffracting crystal unit cell (~1–50 nm in dimension) and the wavelength of the incident radiation (typically 1.54178 Å for Cu Kα and 0.71073 Mo Kα). Thus, x-ray diffraction methods provide a direct way to display the decomposition of x-rays in component waves (frequencies), and for this reason x-ray diffraction may be called *spatial frequency spectrum* analysis or *harmonic* analysis. Near-field and far-field optics are compared in more detail in Chapter 5.

The number of scattered x-ray photons picked up by the detector results in an intensity proportional to the square of the amplitude (peak height) of the diffracted waves. The scattering intensity depends on the number and distribution of electrons associated with the scattering atoms in the basis, i.e., the structure factor $F(hkl)$. The phase of the diffracted rays is the relative time of arrival of the scattered radiation at a particular point in space, say at the emulsion of a photographic film. The phase information is lost when the x-ray diffraction pattern is recorded on the film—one could say that the film integrates intensity over time—in other words, we cannot measure x-ray phases directly. When using a lens such as in a microscope, light first strikes the imaged object and is diffracted in various directions. The lens then collects the diffracted rays and reassembles them to form an image. The *phase problem* is a major concern in structure analysis as we need both intensity and phase to feed into the Fourier transform. Today several techniques exist to regenerate the *lost phase* information of x-ray diffraction, but the topic falls outside the scope of this book. We need to extract the structure factor $F(hkl)$ from the intensities and phase of the diffraction spots and then do an inverse Fourier transform (T^{-1}) to obtain the crystal structure/electron density function, $\rho(x, y, z)$ (see Figure 2.17).

We now mathematically derive the intensity profile of x-rays scattered from a crystal. The result, as Laue predicted, is the same as for visible light diffracted from an optical grating. When an incident

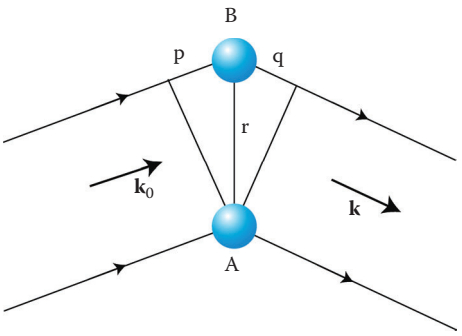


FIGURE 2.27 Scattering of x-rays from two nearby atoms A and B with identical scattering density. (Drawing by Mr. Chengwu Deng.)

x-ray beam travels inside a crystal, we assume that the beam is not much influenced by the presence of the crystal; in other words, the refractive index for x-rays is close to unity, and there is not much loss of energy from the beam through scattering, i.e., elastic scattering dominates!

With reference to Figure 2.27, we assume two parallel plane x-ray waves of wavelength λ and frequency ν (hence velocity $c = \lambda\nu$), scattered elastically from two nearby atoms A and B of identical scattering density. The wave vector for the incoming wave is $\mathbf{k}_0$ and that of the diffracted beam is $\mathbf{k}$. Because we assume elastic scattering:

$$|\mathbf{k}_0| = |\mathbf{k}| = k = 2\pi/\lambda \tag{2.31}$$

Scattering atom A is at the origin, and scattering atom B is at a distance $\mathbf{r}$ away from the origin. The path difference, or phase factor, between the waves can be calculated from Figure 2.27 as:

$$p + q = \mathbf{r}\cdot(\mathbf{k} - \mathbf{k}_0) = \mathbf{r}\cdot\Delta\mathbf{k} \tag{2.32}$$

The equations for the wave amplitudes are (for those readers less familiar with wave equations, consult Appendix 2A for more details):

$$\psi_A(x,t) = \frac{Af}{l_A} e^{i(\mathbf{k}_0\cdot\mathbf{l}_A - \omega\cdot t)} \tag{2.33}$$

and:

$$\psi_B(x,t) = \frac{Af}{l_B} e^{i(\mathbf{k}_0\cdot\mathbf{l}_B - \omega t + \mathbf{r}\cdot\Delta k)} \tag{2.34}$$

where A is the amplitude and f is the atomic scattering factor defined as the ratio of the amplitude of an electromagnetic wave scattered by atom A and that

of a wave scattered by a free electron. The position of the detector with respect to atom A is given by l_A, and the position of the detector relative to atom B is l_B. The product $\mathbf{r}\cdot\Delta\mathbf{k}$ is the phase factor we calculated in Equation 2.32. Generalizing, an x-ray wave scattered from the j_{th} atom in a crystal is:

$$\Psi_j(x,t) = \frac{Af_j}{l_j} e^{i(\mathbf{k}_0\cdot\mathbf{l}_j - \omega t + \mathbf{r}_j\cdot\Delta\mathbf{k})} \tag{2.35}$$

where $\mathbf{r}_j$ is the position of scatterer j relative to scatterer A, and l_j is the position of the detector with respect to scatterer j. The total scattered wave amplitude at the detector is the sum of all the contributing atoms:

$$\begin{aligned}\Psi(x,t) &= \sum_{\text{all atoms}} \frac{Af_j}{l_j} e^{i(\mathbf{k}_0\cdot\mathbf{l}_j - \omega t)} e^{i(\mathbf{r}_j\cdot\Delta\mathbf{k})} \\ &\cong \frac{A}{L} e^{i(\mathbf{k}_0\cdot\mathbf{L} - \omega t)} \sum_{\text{all atoms}} f_j e^{i(\mathbf{r}_j\cdot\Delta\mathbf{k})}\end{aligned} \tag{2.36}$$

For a small sample, the distances l_j are all about the same (the crystal is small compared to the distance between it and the detector), so in Equation 2.36 we can replace l_j with L.

Constructive and destructive interference between the scattered waves that reach the detector is the result of the sum of all scatterers and the scattering vector, $\Delta\mathbf{k}$, and determines where the detector should be put. The term $F(hkl)$:

$$F(hkl) = \sum_{\text{all atoms}} f_j e^{i(\mathbf{r}_j\cdot\Delta\mathbf{k})} \tag{2.37}$$

from the expression 2.36 for the wave amplitude, is the geometrical structure factor. It is defined, in analogy with the atomic scattering factor, as the ratio of the amplitude of the wave scattered by all atoms in a unit cell and that scattered by a free electron for the same incident beam. It incorporates the scattering of all atoms of the unit cell and sums up the extent to which interference of the waves scattered from atoms within the basis diminishes the intensity of the diffraction peaks. So even if the Laue condition is met, the structure factor may be zero and no diffraction will be observed. In the case that there is only one type of atom in the basis, the atomic scattering factor f_j disappears ($f_j = 1$)—identical atoms have identical scattering factors.

Assume now a crystal with base vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ and a total number of atoms along each axis of M, N, and P, respectively, and also accept that there is only a single atom at each lattice point (i.e., $f_j = 1$), then the amplitude of the total wave will be proportional to:

$$\begin{aligned}\Psi(x,t) \propto F(h,k,l) &= \sum_{\text{all atoms}} e^{i(\mathbf{r}_j\cdot\Delta\mathbf{k})} \\ &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{p=0}^{P-1} e^{i[(m\mathbf{a}_1 + n\mathbf{a}_2 + p\mathbf{a}_3)\cdot\Delta\mathbf{k}]}\end{aligned}$$

after rearrangement we obtain:

$$\Psi(x,t) \propto F(h,k,l) = \sum_{m=0}^{M-1} e^{im\mathbf{a}_1\cdot\Delta\mathbf{k}} \sum_{n=0}^{N-1} e^{in\mathbf{a}_2\cdot\Delta\mathbf{k}} \sum_{p=0}^{P-1} e^{ip\mathbf{a}_3\cdot\Delta\mathbf{k}} \tag{2.38}$$

The intensity of the scattered wave is the square of the wave amplitude, and taking the value of one of the sums in Equation 2.38 for a crystal of dimension $M\mathbf{a}_1$ in the direction of $\mathbf{a}_1$, we obtain:

$$\sum_{m=0}^{M-1} e^{im\mathbf{a}_1\cdot\Delta\mathbf{k}} = \frac{1 - e^{iM(\mathbf{a}_1\cdot\Delta\mathbf{k})}}{1 - e^{i(\mathbf{a}_1\cdot\Delta\mathbf{k})}}$$

$$\text{and } I \propto |\Psi(x,t)|^2 \propto \frac{[1 - e^{-iM(\mathbf{a}_1\cdot\Delta\mathbf{k})}][1 - e^{iM(\mathbf{a}_1\cdot\Delta\mathbf{k})}]}{[1 - e^{-i(\mathbf{a}_1\cdot\Delta\mathbf{k})}][1 - e^{i(\mathbf{a}_1\cdot\Delta\mathbf{k})}]}$$

which simplifies to

$$I \propto \frac{2 - e^{-iM(\mathbf{a}_1\cdot\Delta\mathbf{k})} - e^{iM(\mathbf{a}_1\cdot\Delta\mathbf{k})}]}{[2 - e^{-i(\mathbf{a}_1\cdot\Delta\mathbf{k})} - e^{i(\mathbf{a}_1\cdot\Delta\mathbf{k})}]} = \frac{2 - 2\cos(M\mathbf{a}_1\cdot\Delta\mathbf{k})}{2 - 2\cos(\mathbf{a}_1\cdot\Delta\mathbf{k})}$$

$$I \propto \frac{1 - \cos(M\mathbf{a}_1\cdot\Delta\mathbf{k})}{1 - \cos(\mathbf{a}_1\cdot\Delta\mathbf{k})} \quad \text{and with the identity}$$

$$\cos 2x = \cos^2 x - \sin^2 x = 1 - 2\sin^2 x$$

$$I \propto \frac{\sin^2\left(\frac{1}{2}M\mathbf{a}_1\cdot\Delta\mathbf{k}\right)}{\sin^2\left(\frac{1}{2}\mathbf{a}_1\cdot\Delta\mathbf{k}\right)} \tag{2.39}$$

This is the same result as the light intensity profile expected from an M-slit diffraction grating (see Chapter 5 on photonics). If M is large ($\sim 10^8$ for a macroscopic crystal), it has very narrow, intense peaks. Between the peaks the intensity is essentially zero. In Figure 2.28 we have plotted $y = \sin^2 Mx/\sin^2 x$.

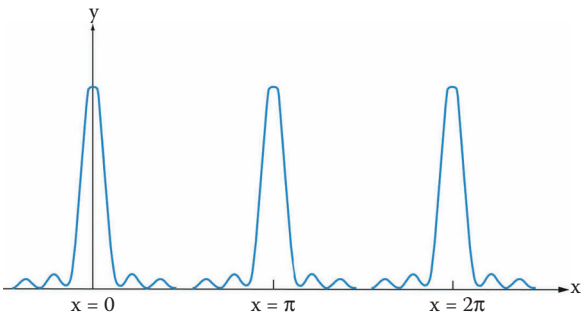


FIGURE 2.28 Graphical presentation of $y = \sin^2 Mx/\sin^2 x$. The width of the peaks and the prominence of the ripples are inversely proportional to *M*.

This function is virtually zero except at the points where $x = n\pi$ (*n* is an integer including zero), where it rises to the maximum value of M^2. The width of the peaks and the prominence of the ripples are inversely proportional to *M*.

Remember that there are three sums in Equation 2.38. For simplicity we only evaluated one sum to calculate the intensity in Equation 2.39. The total intensity equals:

$$I \propto \frac{\sin^2\left(\frac{1}{2}M\mathbf{a}_1 \cdot \Delta\mathbf{k}\right)}{\sin^2\left(\frac{1}{2}\mathbf{a}_1 \cdot \Delta\mathbf{k}\right)} \times \frac{\sin^2\left(\frac{1}{2}N\mathbf{a}_2 \cdot \Delta\mathbf{k}\right)}{\sin^2\left(\frac{1}{2}\mathbf{a}_2 \cdot \Delta\mathbf{k}\right)} \times \frac{\sin^2\left(\frac{1}{2}P\mathbf{a}_3 \cdot \Delta\mathbf{k}\right)}{\sin^2\left(\frac{1}{2}\mathbf{a}_3 \cdot \Delta\mathbf{k}\right)} \tag{2.40}$$

so that the diffracted intensity will equal zero unless all three quotients in Equation 2.40 take on their maximum values at the same time. This means that the three arguments of the sine terms in the denominators must be simultaneously equal to integer multiples of 2π, or the peaks occur only when:

$$\mathbf{a}_1 \cdot \Delta\mathbf{k} = 2\pi e$$

$$\mathbf{a}_2 \cdot \Delta\mathbf{k} = 2\pi f$$

$$\mathbf{a}_3 \cdot \Delta\mathbf{k} = 2\pi g$$

These are, of course, the familiar Laue equations. As we will see below, to solve for the wave vector $\Delta\mathbf{k}$ it is very convenient to introduce the concept of a reciprocal lattice so that the set of all wave vectors $\Delta\mathbf{k}$ yields plane waves with the periodicity of a given Bravais lattice.

From the above, the intensity of the x-ray diffraction is:

$$I \propto |F(X)|^2 = \left|\sum_{\text{all atoms}} f_j e^{i(\mathbf{r}_j \cdot \Delta\mathbf{k})}\right|^2 \tag{2.41}$$

So what we measure from the x-ray film is intensity *I*, and the dependence of that intensity on the atomic position follows from the Fourier expansion of the electron density function in Equation 2.17:

$$\rho(x,y,z) = \sum_h \sum_k \sum_l F(hkl) e^{i(\mathbf{r}_j \cdot \Delta\mathbf{k})} \tag{2.42}$$

where $F(hkl)$ is the coefficient to be determined, and *h*, *k*, and *l* are integers over which the series is summed. Because of the 3D periodicity, a triple summation is required:

$$F(hkl) = \int_0^1 \int_0^1 \int_0^1 \rho(x,y,z) e^{i(\mathbf{r}_j \cdot \Delta\mathbf{k})} dx dy dz \tag{2.43}$$

If we knew all the $F(hkl)$ values, we could calculate the electron density, and vice versa. Unfortunately, as remarked above, Equation 2.42 requires values of $F(hkl)$, but the measured intensities only give us $|F(hkl)|^2$. The apparent impasse is known as the *phase problem* and arises because we need to know both the amplitude and the phase of the diffracted waves to compute the inverse Fourier transform.

Reciprocal Space, Fourier Space, k-Space, or Momentum Space

Introduction

The crystal lattice discussed above is a lattice in real space. Practice has shown the usefulness of defining a lattice in reciprocal space for the interpretation of diffraction patterns of single crystals, the simple reason being that a diffraction pattern reveals the reciprocal lattice. We did already consider some reciprocal features when we introduced the Miller indices, which are derived as the reciprocal of the unit cell intercepts. Because the lattice distances

between planes are proportional to the reciprocal of the distances in the real crystal (i.e., they are proportional to $1/d_{hkl}$), the array is called a reciprocal lattice. We will learn in this section that the condition for nonzero intensity in scattered x-rays is that the scattering vector $\Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$ is a translation vector, $\mathbf{G}_{hkl}$, of the reciprocal lattice, i.e., $\Delta\mathbf{k} = \mathbf{G}_{hkl}$. One can describe a reciprocal lattice the same way we describe a real one, but one must keep in mind that in one case the points are the position of real objects (atoms or the base), whereas in the other case they mark the positions of abstract points—magnitude and direction of momentum. Reciprocal space, also called Fourier space, **k**-space, or momentum space, plays a fundamental role in most analytical studies of periodic structures. After the introduction of the reciprocal lattice we will discuss the conditions for x-ray diffraction in terms of such a lattice and end with an introduction to Brillouin zones. The concepts introduced here are not only needed to understand x-ray diffraction better, they also are required preparation for Chapter 3, where we deal with band theory of solids.

The Reciprocal Lattice

Graphical Presentation of the Reciprocal Lattice

Before introducing a reciprocal lattice mathematically, let us learn how to draw one as we demonstrate in Figure 2.29. In this example we start with the two-dimensional unit cell of a monoclinic crystal with $\mathbf{a}_1$ and $\mathbf{a}_2$ in the plane of the paper. The figure also shows the edges of four (*hkl*) planes: (100), (110), (120), and (010). These planes all are perpendicular to the face of the paper. To construct the reciprocal lattice we draw from a common origin a normal to each plane. Next we place a point on the normal to each plane (*hkl*) at a distance $2\pi/d_{hkl}$ from the origin. Each of the points thus obtained maintains the important features of the stack of parallel planes it represents. The direction of the point from the origin preserves the orientation of the planes, and the distance of the point from the origin preserves the interplanar distance. A doubling of the periodicity in real space will produce twice as many diffraction spots in reciprocal space. It is convenient to let the reciprocal lattice vector, $\mathbf{G}_{hkl}$, be 2π times the reciprocal of the

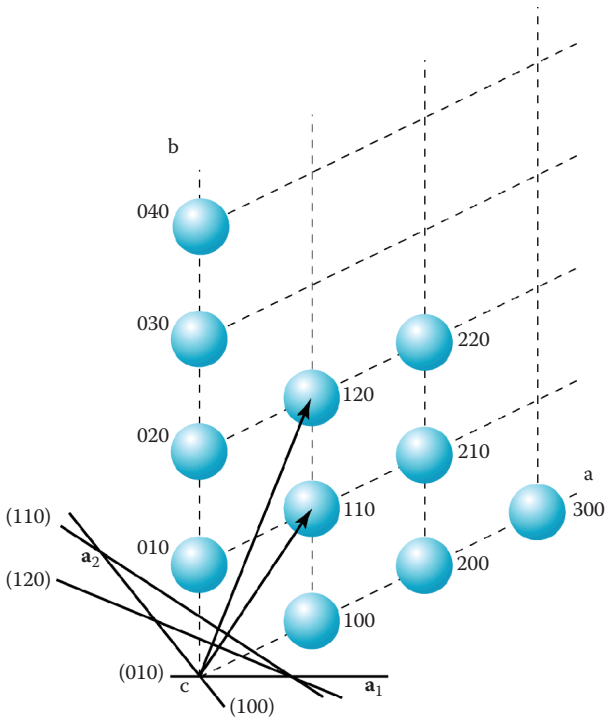


FIGURE 2.29 Graphical construction of the reciprocal lattice. (Drawing by Mr. Chengwu Deng.)

interplanar distance, d_{hkl}. This convention converts the units from periods per unit length to radians per unit length. This simplifies comparison of different periodic phenomena, e.g., a crystal lattice and light that interacts with it. For instance, $\Delta\mathbf{k}$, the wave vector, has an absolute value of $2\pi/\lambda$. If we choose this scaling factor, we are able to compare the two values directly, and all it really does in our drawing is to expand the size of the reciprocal lattice.

Definition of the Reciprocal Lattice

Now that we can draw a reciprocal lattice in 2D, let us define what a reciprocal lattice is. Imagine a set of points **R** constituting a Bravais lattice and a plane wave, $e^{i\mathbf{k}\cdot\mathbf{r}}$, interacting with that lattice. For most values of **k**, a plane wave will not have the periodicity of the Bravais lattice of points **R**. Only a very special set of the wave vectors, $\mathbf{G}_{hkl} = h\cdot\mathbf{a}_1^* + k\cdot\mathbf{a}_2^* + l\cdot\mathbf{a}_3^*$, of the reciprocal lattice satisfies that condition. Mathematically, $\mathbf{G}_{hkl}$ belongs to the reciprocal lattice of points **R** of the Bravais lattice, if the relation:

$$e^{i\mathbf{G}_{hkl}\cdot(\mathbf{r}+\mathbf{R})} = e^{i\mathbf{G}_{hkl}\cdot\mathbf{r}} \qquad (2.44)$$

holds for any **r**, and for all **R** in the Bravais lattice. From Equation 2.44 it follows that we can describe

the reciprocal lattice as the set of wave vectors $\mathbf{G}_{hkl}$ satisfying:

$$e^{i\mathbf{G}_{hkl}\cdot\mathbf{R}} = 1 \tag{2.45}$$

for all $\mathbf{R}$ in the Bravais lattice. Because lattice scatterers are displaced from one another by the lattice vectors $\mathbf{R}$, the condition that all scattered rays interfere constructively is that the Laue equations hold simultaneously for all values of Bravais lattice vectors $\mathbf{R}$, or:

$$\mathbf{R}\cdot\Delta\mathbf{k} = 2\pi m \tag{2.46}$$

with integer m. Because $\exp[i2\pi(\text{integer})] = 1$, this can be written in the equivalent form:

$$ei^{\mathbf{R}}\cdot\Delta^{\mathbf{k}} = 1 \tag{2.47}$$

We see from Equation 2.47 that if $\Delta\mathbf{k}$ is equal to any reciprocal lattice vector $\mathbf{G}_{hkl}$, then the Laue equations for wave diffraction are satisfied. The diffraction condition is thus simply:

$$\Delta\mathbf{k} = \mathbf{G}_{hkl} \tag{2.48}$$

or from Equation 2.23 (with e = h, f = k and g = l):

$$\begin{aligned}\mathbf{G}_{hkl}\cdot\mathbf{a}_1 &= 2\pi h\\ \mathbf{G}_{hkl}\cdot\mathbf{a}_2 &= 2\pi k\\ \mathbf{G}_{hkl}\cdot\mathbf{a}_3 &= 2\pi l\end{aligned} \tag{2.49}$$

The expression $\Delta\mathbf{k} = \mathbf{G}_{hkl}$ can be represented in the form of a vector triangle as illustrated in Figure 2.30. Because $\mathbf{k}$ and $\mathbf{k}_0$ are of equal length, i.e. $2\pi/\lambda$, the triangle O, O', O'' has two equal sides. The angle between $\mathbf{k}$ and $\mathbf{k}_0$ is 2θ, and the *hkl* plane dissects it.

In the two-dimensional reciprocal lattice shown at the bottom of Figure 2.31, the $\mathbf{G}_{hkl}$ vectors give the outline of the unit cell. Vector $\mathbf{a}_1^* = \mathbf{G}_{100}$ and $\mathbf{a}_3^* = \mathbf{G}_{001}$ and:

$$|\mathbf{a}_1^*| = \frac{2\pi}{d_{100}} \text{ and } |\mathbf{a}_3^*| = \frac{2\pi}{d_{001}} \tag{2.50}$$

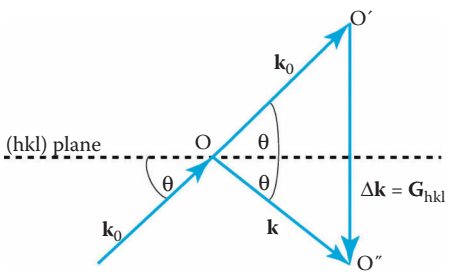


FIGURE 2.30 Vector triangle representation of $\Delta\mathbf{k} = \mathbf{G}_{hkl}$. (Drawing by Mr. Chengwu Deng.)

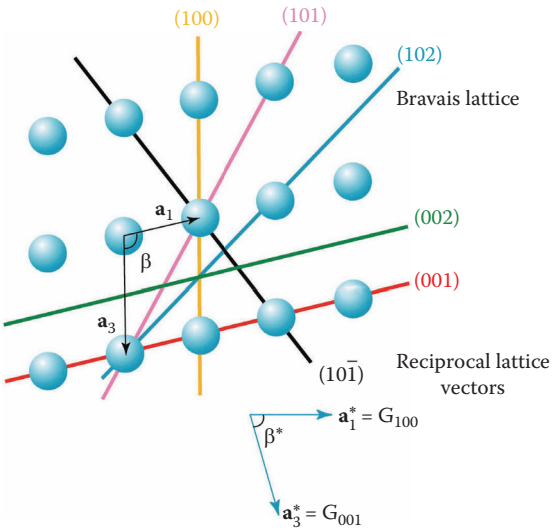


FIGURE 2.31 Reciprocal lattice vectors $a_1^* = G_{100}$ and $a_3^* = G_{001}$ in a monoclinic unit cell and their relation to the Bravais lattice. (Drawing by Mr. Chengwu Deng.)

The angle β^* between the reciprocal lattice vectors in Figure 2.31 is the complement of β in the Bravais lattice.

The components of any vector referred to the reciprocal lattice represent the Miller indices of a plane whose normal is along that vector, whereas the spacing of the plane is given by the inverse of the magnitude of that vector multiplied by 2π. For example, the reciprocal lattice vector $\mathbf{u}^* = [123]$ is normal to the planes with Miller indices (123) and has an interplanar spacing $2\pi/|\mathbf{u}^*|$.

Explicit Algorithm for Constructing the Reciprocal Lattice

We can write out the change in wave vector, $\mathbf{G}_{hkl} = \Delta\mathbf{k} = \mathbf{k} - \mathbf{k}_0$, in the following expression:

$$\mathbf{G}_{hkl} = \Delta\mathbf{k} = e\mathbf{a}_1^* + f\mathbf{a}_2^* + g\mathbf{a}_3^* \tag{2.51}$$

where *e*, *f*, and *g* are the integers from the Laue equations (Equation 2.23) and $\mathbf{a}_1^*$, $\mathbf{a}_2^*$ and $\mathbf{a}_3^*$ are basis vectors of the reciprocal lattice to be determined. Equation 2.51 is a solution of the Laue Equation 2.23 if all of the following relations are satisfied:

$$\begin{aligned}&\mathbf{a}_1^*\cdot\mathbf{a}_1 = \mathbf{a}_2^*\cdot\mathbf{a}_2 = \mathbf{a}_3^*\cdot\mathbf{a}_3 = 2\pi\\ &\mathbf{a}_1^*\cdot\mathbf{a}_2 = \mathbf{a}_2^*\cdot\mathbf{a}_1 = \mathbf{a}_3^*\cdot\mathbf{a}_2\\ &= \mathbf{a}_1^*\cdot\mathbf{a}_3 = \mathbf{a}_2^*\cdot\mathbf{a}_3 = \mathbf{a}_3^*\cdot\mathbf{a}_1 = 0\end{aligned} \tag{2.52}$$

From Equation 2.52, $\mathbf{a}_1^*$ is perpendicular to primitive lattice vectors $\mathbf{a}_2$ and $\mathbf{a}_3$ of the direct lattice. A vector

TABLE 2.3 Bravais Lattice after Fourier Transform

Real Space	Fourier Space
Normals to the planes (vectors)	Points
Spacing between planes	2π/distance planes
λ is distance, wavelength	$2\pi/\lambda$ is momentum or wave number
First Brillouin zone (see below)	Wigner-Seitz cell

that is perpendicular to $\mathbf{a}_2$ and $\mathbf{a}_3$ is given by the vector product $\mathbf{a}_2 \times \mathbf{a}_3$. To construct $\mathbf{a}_1^*$ of the reciprocal lattice completely, we must normalize $\mathbf{a}_2 \times \mathbf{a}_3$ to satisfy the expression $\mathbf{a}_1^* \cdot \mathbf{a}_1 = 2\pi$. All the equations in 2.52 are thus satisfied when choosing:

$$\mathbf{a}_1^* \equiv \frac{2\pi \mathbf{a}_2 \times \mathbf{a}_3}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}$$
$$\mathbf{a}_2^* \equiv \frac{2\pi \mathbf{a}_3 \times \mathbf{a}_1}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3}$$
$$\mathbf{a}_3^* \equiv \frac{2\pi \mathbf{a}_1 \times \mathbf{a}_2}{\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3} \qquad (2.53)$$

The denominators have been written the same way because of the property of the triple scalar product: $\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3 = \mathbf{a}_2 \cdot \mathbf{a}_3 \times \mathbf{a}_1 = \mathbf{a}_3 \cdot \mathbf{a}_1 \times \mathbf{a}_2$. The magnitude of this triple product is the volume of the primitive cell. In Table 2.3 we summarize the properties of a Bravais lattice before and after a Fourier transform.

The Ewald Construction

The Ewald construction is a way to visualize the Laue condition that the change in wave vector, $\Delta\mathbf{k}$, must be a vector of the reciprocal lattice. We reconsider for a moment the vector representation of the von Laue condition $\Delta\mathbf{k} = \mathbf{G}_{hkl}$ represented in the form of a vector triangle in Figure 2.30. Because $\mathbf{k}$ and $\mathbf{k}_0$ are of equal length ($= 2\pi/\lambda$) the triangle O, O', O'' has two equal sides, and we can draw a sphere with $|\mathbf{k}| = |\mathbf{k}_0| = 2\pi/\lambda$ as illustrated in Figure 2.32. The angle between $\mathbf{k}$ and $\mathbf{k}_0$ is 2θ, and the *hkl* plane dissects it. Diffraction of the incoming beam represented by the vector $\mathbf{k}_0$ giving the vector $\mathbf{k}$ may be thought of as reflection from the dotted line in this diagram (as in Bragg's Law). We will now see that with $\Delta\mathbf{k} = \mathbf{G}_{hkl}$ the diffraction vector corresponds to the distance between planes in reciprocal space. We thus superimpose an imaginary sphere of x-ray radiation on the reciprocal lattice as illustrated

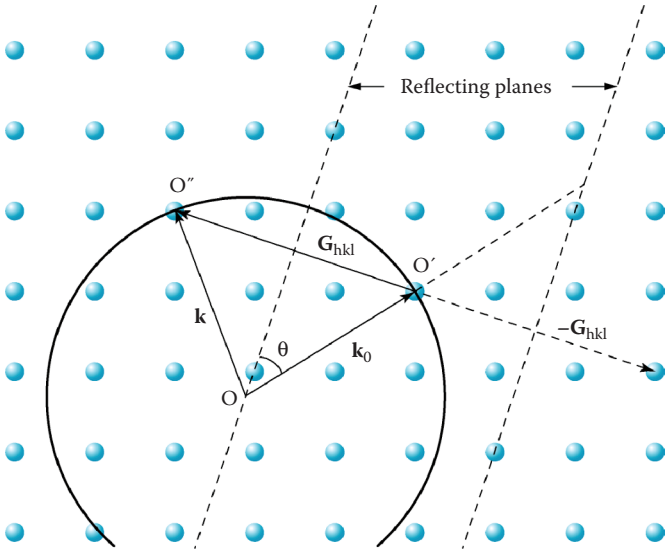


FIGURE 2.32 The reciprocal lattice and the geometry of diffraction clarified by the Ewald sphere. The sphere with center *O* intersects the reciprocal lattice center *O′*. (Drawing by Mr. Chengwu Deng.)

in Figure 2.32. Somewhat confusingly, one might consider two origins: *O*, which is the center of the sphere and may be considered as the position of the crystal, is the source of the secondary beam $\mathbf{k}$, and *O′* is the origin of reciprocal space, the origin of the diffraction vector $\Delta\mathbf{k} = \mathbf{G}_{hkl}$, and the center of the reciprocal lattice. As the crystal rotates, the reciprocal lattice rotates in exactly the same way. Any points in reciprocal space that intersect the surface of the sphere reveals where diffraction peaks will be observed if the structure factor is nonzero. In other words, reflection is only observed if the sphere intersects a point where $\Delta\mathbf{k} = \mathbf{G}_{hkl}$. The diffraction angle, θ, is then half the angle between the incident and diffracted wave vectors.

It is useful to think of the crystal at the center of the Ewald sphere (*O*) being linked to the center (origin) of the reciprocal lattice (*O′*) by something like a bicycle chain—the two "objects" rotate exactly in step with each other.

The Ewald construction also makes for a good link with the Brillouin zones discussed in the next section. Notice in Figure 2.32 that both $\mathbf{G}_{hkl}$ and $-\mathbf{G}_{hkl}$ are vectors in reciprocal space. It can be seen that a reflecting plane bisects the vector $\mathbf{G}_{hkl}$ at $\mathbf{G}_{hkl}/2$ and another reflecting plane cuts $-\mathbf{G}_{hkl}$ at $-\mathbf{G}_{hkl}/2$. The incident wave vector $\mathbf{k}_0$ starting from the Ewald circle's center *O* must terminate at the $-\mathbf{G}_{hkl}/2$ reflecting plane for

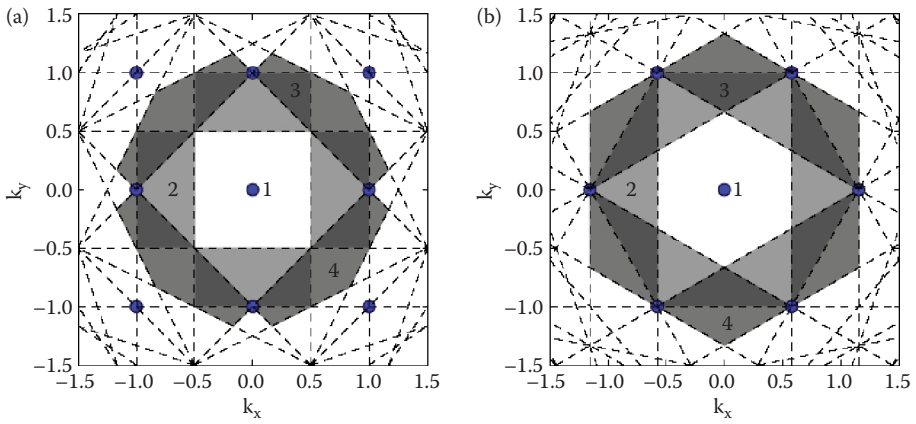


FIGURE 2.33 The Brillouin zones for (a) a square 2D lattice and (b) a triangular 2D lattice. The solid circles are the lattice points, and the dashed lines are the Bragg lines. The first four Brillouin zones are marked with different gray scales.

diffraction to occur. This observation gives rise to the idea of using the reciprocal lattice to construct Brillouin zones, the boundaries of which satisfy the Laue conditions for diffraction. Bragg planes bisect the lines joining the origin to points of the reciprocal lattice, and one can define the first Brillouin zone as the set of points that can be reached from the origin without crossing any other Bragg planes. Recall that this was also the way we learned how to construct Wigner-Seitz cells. So the first Brillouin zone is the Wigner-Seitz cell of the reciprocal lattice.

Brillouin Zones

Leon Brillouin (1889–1969) was a French-American physicist who was a professor at the Sorbonne, Collège de France, Wisconsin, Columbia, and Harvard. He also worked briefly at IBM. In the previous section we used the Ewald sphere (Figure 2.32) as a way to introduce the important concept of Brillouin zones. A Brillouin zone is defined as a Wigner-Seitz cell in the reciprocal lattice and gives a geometric interpretation of the diffraction condition. The Wigner-Seitz cell of the reciprocal lattice is the set of points laying closer to $\Delta\mathbf{k} = 0$ than to any other reciprocal lattice point. The Brillouin construction exhibits all wave vectors $\mathbf{k}$ that can be Bragg reflected by the crystal. The constructions divide the Fourier space into zones, out of which the first Brillouin zone is of greatest importance.

We have just seen that x-ray waves traveling in a crystal lattice undergo Bragg reflection at certain wave vectors. The first Brillouin zone is the set of points in $\mathbf{k}$-space that can be reached from the origin without crossing any Bragg reflection plane. The second Brillouin zone is the set of points that can be reached from the first zone by crossing only one Bragg plane. The nth Brillouin zone can be defined as the set of points that can be reached from the origin by crossing $n - 1$ Bragg planes, but no fewer. Each Brillouin zone is a primitive cell of the reciprocal lattice. In Figure 2.33 we illustrate the first four Brillouin zones in square and triangular two-dimensional Bravais lattices. For examples of 3D Wigner-Seitz cells, corresponding to Brillouin zones, see Figures 2.6 and 2.34.

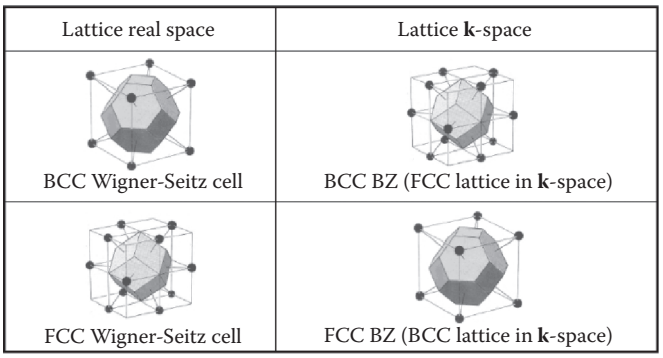


FIGURE 2.34 The Wigner-Seitz cell of BCC lattice in real space transforms to a Brillouin zone in an FCC lattice in reciprocal space, whereas the Wigner-Seitz cell of an FCC lattice transforms to a Brillouin zone of a BCC lattice in reciprocal space.

Nothing Is Perfect

Even a crystal having a mass of less than 0.1 g is likely to have more than 10^{23} ions, and it is unrealistic to

expect that the growth of such crystal from solution can lead to zero impurities or imperfections. Successive purification steps may remove impurities, but even in the purest crystal, thermodynamics predicts the existence of some structural intrinsic imperfections. There are at least four types of defects one distinguishes: 1) point defects where the irregularity in structure extends over only a few atoms in size (0D), 2) linear defects where the irregularity extends across a single row of atoms (1D), 3) planar defects with an irregularity extending across a plane of atoms (2D), and 4) volumetric defects with the irregularity taking place over 3D clusters of atoms. In addition, defects can be categorized as either intrinsic, where defects are induced because of physical laws, and extrinsic, where defects are present because of the environment and/or processing conditions.

A physical law imposing the presence of intrinsic defects in a crystal is the minimization of the Gibbs free energy (G). The Gibbs free energy is higher in a crystal without vacancies than one with vacancies (G = H – TS = minimum). This is because defects are energetically unfavorable but are entropically favorable. Formation of vacancies (n) does indeed increase the enthalpy H of the crystal because of the energy required to break bonds ($\Delta H = n\Delta H_f$), but vacancies also increase S of the crystal as a result of an increase in configurational entropy. The configurational entropy is given as:

$$S = k\ln W \tag{2.54}$$

where W is the number of microstates. If the number of atoms in the crystal is N and the number of vacancies is n, then the total number of sites is $n + N$, and the number of all possible microstates W may be calculated from:

$$W = \frac{(N+n)!}{n!N!} \tag{2.55}$$

and the increase in entropy ΔS is then given by:

$$\Delta S = k\ln W = k\ln\frac{(N+n)!}{n!N!} \tag{2.56}$$

and the total free energy change as:

$$\Delta G = n\Delta H_f - T\Delta S \tag{2.57}$$

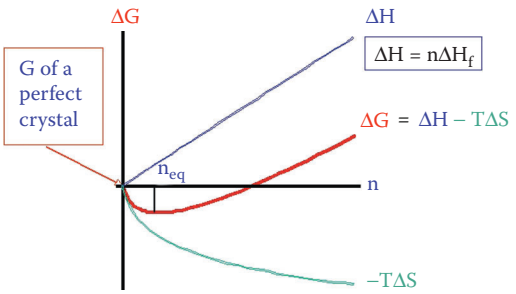


FIGURE 2.35 Change in Gibbs free energy G of a crystal as a result of the number of vacancies n.

This expression is plotted in Figure 2.35. It can be seen that for a crystal in equilibrium, vacancies are required to be present at any temperature above 0 K. The equilibrium concentration n_{eq} is calculated from:

$$\left|\frac{\partial \Delta G}{\partial n}\right|_{n=n_{eq}} = 0 \tag{2.58}$$

or we obtain:

$$\frac{n_{eq}}{N} = \exp\left(-\frac{\Delta H_f}{kT}\right) \tag{2.59}$$

so the n_{eq} should increase with increasing temperature. For Al, $\Delta H_f = 0.70$ eV/vacancy and for Ni, $\Delta H_f = 1.74$ eV/vacancy, leading to the values for n_{eq}/N at three different temperatures as shown in Table 2.4.

Defects, even in very small concentrations, can have a dramatic impact on the properties of a material. Actually, without defects solid-state electronic devices could not exist, metals would be much stronger, ceramics would be much tougher, and some crystals would have no color. Vacancies even make a small contribution to the thermal expansion of a crystal. Some commonly observed defects are summarized in Table 2.5; they are categorized here as point defects (0D), line defects (1D), and plane defects (2D).

The simplest sorts of defects are those based on points (0D). In Figure 2.36 we review point defects, including vacancies, interstitial atoms, small and

TABLE 2.4 n_{eq}/N Values for Al and Ni at Three Different Temperatures

n_{eq}/N	0 K	300 K	900 K
Al	0	$1.45\ 10^{-12}$	$1.12\ 10^{-4}$
Ni	0	$5.59\ 10^{-30}$	$1.78\ 10^{-10}$

Table 2.5 Common Defects in Crystals

Type of Imperfection	Description
Point defects:	
Interstitial	Extra atom in an interstitial site
Schottky defect	Atom missing from correct site
Frenkel defect	Atom displaced to interstitial site creating nearby vacancy
Line defects:	
Edge dislocations	Row of atoms marking edge of a crystallographic plane extending only part way in crystal
Screw dislocations	Row of atoms about which a normal crystallographic plane appears to be spiral
Plane defects:	
Lineage boundary	Boundary between two adjacent perfect regions in the same crystal that are slightly tilted with respect to each other
Grain boundary	Boundary between two crystals in a polycrystalline solid
Stacking fault	Boundary between two parts of a closest packing having alternate stacking sequences

large substitutional atoms, and the intrinsic Frenkel and Schottky defects.

A missing atom is a vacancy; a dissimilar atom in a nonlattice spot is an interstitial; and a different atom in a lattice position is substitutional. Vacancies are required in ionic solids, just like they are in metals, but the vacancies must be formed in such a way that the solid remains charge neutral. The two main ways

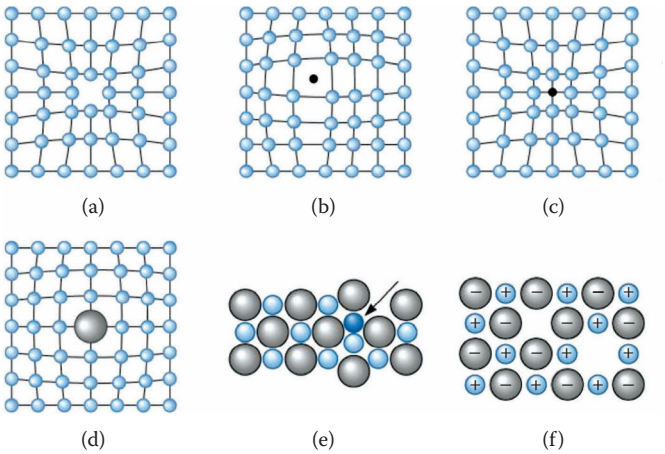


FIGURE 2.36 Point defects: vacancy (a), interstitial atom (b), small (c) and large (d) substitutional atom, Frenkel (e), and Schottky defect (f).

to create point defects in ionic solids without causing a charge imbalance are through correlated vacancies (Schottky defects, for the German who described them in 1930) and through correlated vacancy/interstitial pairs (Frenkel defects, for the Russian who first described them in 1924). Impurities, such as dopants in single-crystal Si (see Chapter 4), are atom(s) of a type that do not belong in the perfect crystal structure. An impurity atom in a pure crystal will generally raise the enthalpy (H) somewhat and increase the entropy (S = klnW) a lot. In other words, there will always be some impurities present for the same reason as vacancies form in a pure crystal. When doping a crystal we introduce impurities on purpose (see Chapter 4), perhaps turning the material into an extrinsic semiconductor. Because of defects, metal oxides may also act as semiconductors. Some nonstoichiometric solids are engineered to be n-type or p-type semiconductors. Nickel oxide (NiO) gains oxygen on heating in air, producing Ni^{3+} sites acting as electron traps, resulting in p-type semiconductor behavior. On the other hand, ZnO loses oxygen on heating, and the excess Zn metal atoms in the sample are ready to donate electrons, leading to n-type semiconductor behavior.

Color centers are imperfections in crystals that cause color. The simplest color center is found in sodium chloride, normally a colorless crystal. When sodium chloride is bombarded with high-energy radiation, a Cl^- can be ejected, creating a vacancy. Momentarily the crystal is no longer electrically neutral, and to regain stability, it grabs an available electron and sticks it in the vacancy previously occupied by the ejected Cl^-. With the electron replacing the ejected Cl^-, there are now equal numbers of positive and negative charges in the crystal, and the electron is held firmly in its site by the surrounding positively charged Na^+ ions. This process turns the colorless salt crystal into an orange/brown. These electrons are color centers, often referred to as F-centers, from the German word *Farben*, meaning color. The color center can also exist in an excited state, and the energy needed to reach that excited state is equal to the energy of a visible photon. The color center absorbs a "violet" photon, causing a jump to the excited state, and the crystal appears with the color orange/brown (the complement of violet). Analogous color

FIGURE 2.37 Color centers in some well-known minerals: (a) the Dresden green diamond, (b) smoky quartz, and (c) amethyst or violet quartz.

centers occur in several minerals. For example, a diamond with C vacancies (missing carbon atoms) absorbs light, and these centers lead to a green color (Figure 2.37). In some cases, impurities are involved in forming the color centers. Replacement of Al^{3+} for Si^{4+} in quartz gives rise to the color of smoky quartz (Figure 2.37). An iron impurity is responsible for the violet color in amethyst through the creation of a color center. The color center, not the iron impurity, is responsible for absorbing the "yellow" photon that makes amethyst violet (Figure 2.37).

Line defects are 1D imperfections in a crystal structure in which a row of atoms has a local structure that differs from the surrounding crystal. These defects are extrinsic because their presence is not required by thermodynamics. They are created by material processing conditions and by mechanical forces acting on the material. In a typical material, about 5 of every 100 million atoms (0.000005%) belong to a line defect. Line defects or dislocations have a dramatic impact on the mechanical properties of metals and some ceramics. Two Hungarians, Egon Orowan and Michael Polanyi, and an Englishman, G.I. Taylor, discovered dislocations in 1934. As we will see, dislocations really govern the mechanical properties of solids, and their discovery was a very important milestone in the material science field. There are two pure types of dislocations, edge dislocations and screw dislocations, but sometimes a mixed type is displayed. An edge dislocation is the simplest type of dislocation and can be viewed as an extra half-plane of atoms inserted into the crystal. This plane terminates somewhere inside the crystal. The boundary of the half-plane is the dislocation, shown in Figure 2.38A. A screw dislocation, shown in Figure 2.38B, is a dislocation produced by skewing a crystal so that one atomic plane produces a spiral ramp about the dislocation. A mixed dislocation is a dislocation that contains some edge components and some screws components as illustrated in Figure 2.38C. Line dislocations cannot terminate inside an otherwise perfect crystal but do end at a crystal surface, an internal surface, or interface (e.g., a grain boundary) or they form dislocation loops. Line defects are mostly caused by the misalignment of ions or the presence of vacancies along a line. When lines of ions are missing in an otherwise perfect array of ions, an edge dislocation appears.

Dislocations are visible in transmission electron microscopy (TEM) where diffraction images of dislocations appear as dark lines. When a crystal surface is etched the rate of material removal at the location of a dislocation is faster, and this results in an etch pit. Symmetrical strain fields of edge dislocations produce a conical pit, whereas for a screw dislocation a spiral etch pit results.

A Burgers vector **b** is a measure of the magnitude and direction of a dislocation. Imagine going around a dislocation line, and exactly going back as many atoms in each direction as you have gone forward, you will not come back to the same atom where you have started. The Burgers vector points from the start atom to the end atom of your journey. This "journey" is called Burgers circuit in dislocation theory. A Burgers circuit is thus a clockwise trace around the core of a dislocation, going from lattice point to lattice point, and it must go an equal number of steps left and right and an equal number

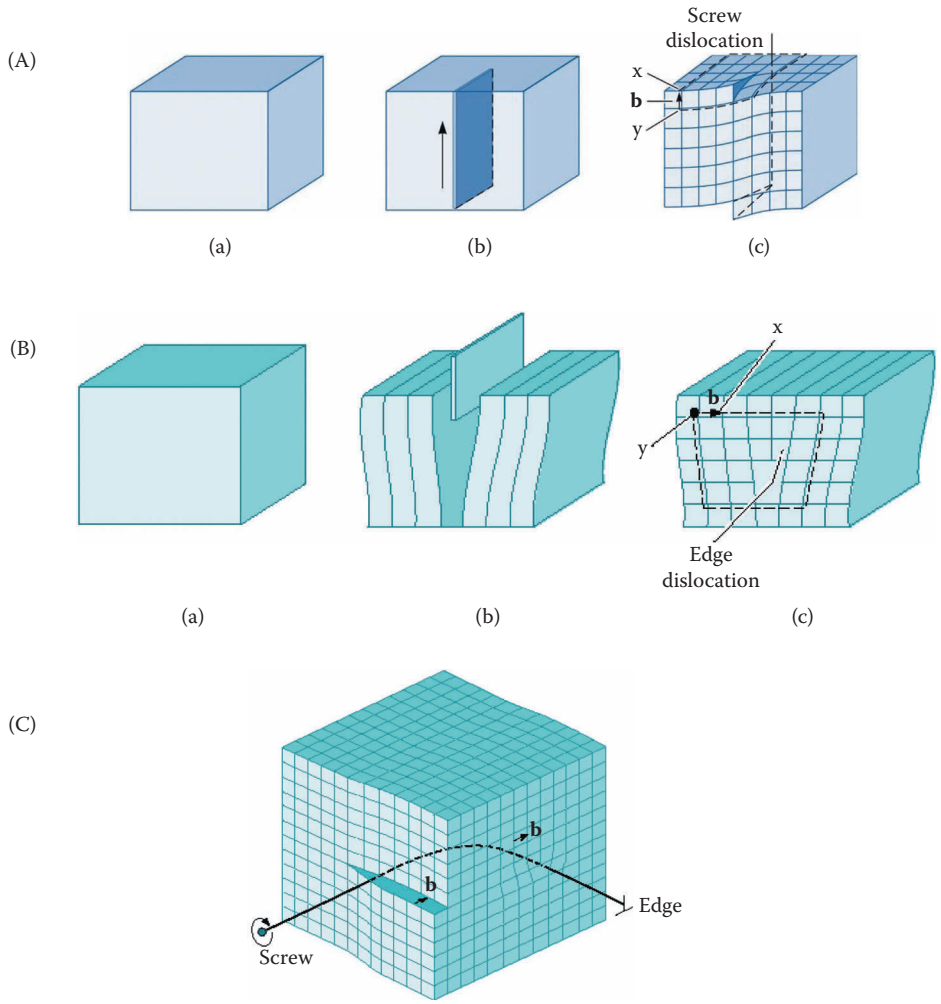


FIGURE 2.38 (A) Edge dislocation: The perfect crystal in (a) is cut, and an extra plane of atoms is inserted (b). The bottom edge of the extra plane is an edge dislocation (c). A Burgers vector **b** is required to close a loop of equal atom spacings around the edge dislocation. (B) Screw dislocation: The perfect crystal (a) is cut and sheared over one atom spacing (b and c). The line along which shearing occurs is a screw dislocation. (C) A mixed dislocation: The screw dislocation at the front face of the crystal gradually changes to an edge dislocation at the side of the crystal.

of steps up and down. Referring to Figure 2.38A, the extra half-plane of lattice points in the edge dislocation causes the Burgers circuit to be open. The vector that points from the end of the Burgers circuit to its beginning is shown here as the Burgers vector, **b**. The vector always points from one lattice point to another; it always has the same length and direction for a given dislocation, regardless where the circuit starts. For an edge dislocation, **b** is always perpendicular to the dislocation line. If **b** is parallel to the dislocation line, the dislocation is a screw dislocation. If **b** is neither perpendicular nor parallel to the line, a mixed dislocation is involved.

Plastic deformation of a material refers to irreversible deformation or change in shape that remains even when the force or stress causing it is removed, whereas elastic deformation is deformation that is fully recovered when the stress causing it is removed. Before the discovery of dislocations, materials scientists faced a big theoretical problem; they calculated that plastic deformation of a perfect crystal should require stresses 100 to 1000 times higher than those observed in tensile tests! So the problem was to explain why metals yielded so easily to plastic deformation. How they got these very large theoretical numbers can be understood from Figure 2.39. Planes of atoms in a crystal slip with respect to each other, in contrast to flow in a fluid, where the solid remains crystalline. The theoretical maximal shear stress or yield strength (failure), τ_{max}, needed

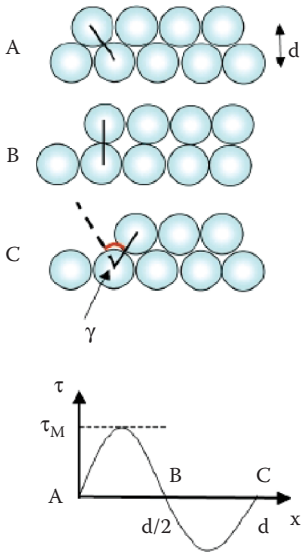


FIGURE 2.39 Calculation of the theoretical shear stress in a crystal.

to produce slip in an ideal crystal is calculated from the type and strength of the bonds involved, the spacing of the crystal, *d*, and the crystal symmetry. In yield, atoms slide tangentially from one equilibrium position to another. Thus, the shear stress, τ, is a periodic function:

$$\tau = \tau_{max} \sin \frac{2\pi x}{d} \qquad (2.60)$$

where *x* is the direction of the shear. With *x* small this may be rewritten as:

$$\tau = \tau_{max} \frac{2\pi x}{d}$$
$$x = \gamma d = \frac{\tau}{\mu} d \therefore \tau = \tau_{max} \frac{2\pi}{d} \frac{\tau}{\mu} d \qquad (2.61)$$
$$\tau_{max} = \frac{\mu}{2\pi}$$

where we have used $\gamma = \tau/\mu$ (Hooke's law) as the shear strain, with μ as the shear modulus (see Figure 2.40).

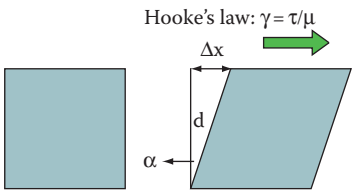


FIGURE 2.40 Shear stress, τ; shear strain, γ; and shear modulus, μ. The shear stress τ produces a displacement Δx of the upper plane as indicated; the shear strain, γ, with $\Delta x/d = \tan \alpha$ is defined $\gamma = \tau/\mu$.

TABLE 2.6 Theoretical Yield Strengths (Shear and Tensile) of Some Important Materials

Material	Shear Strength (GPa)	Tensile Strength (GPa)
Metallic Cu	1.2	3.9
NaCl	2.84	0.43
Quartz	4.4	16
Diamond	121	205

Calculated values for μ are in the range of 1–150 GPa (see also Table 2.6).

But when the experiment is carried out, permanent deformation takes place by a stress as low as 0.5 MPa! The reason is that real materials have lots of dislocations, from $10^2/cm^2$ in the best Ge and Si crystals to $10^{12}/cm^2$ in heavily deformed metals. Therefore, the strength of the material depends on the force required to make dislocations move, not the bonding energy between all atoms in two planes as calculated above. Dislocations can move if the atoms from surrounding planes break their bonds and rebond with the atoms at the terminating edge as shown in Figure 2.41. Instead of all the atoms in

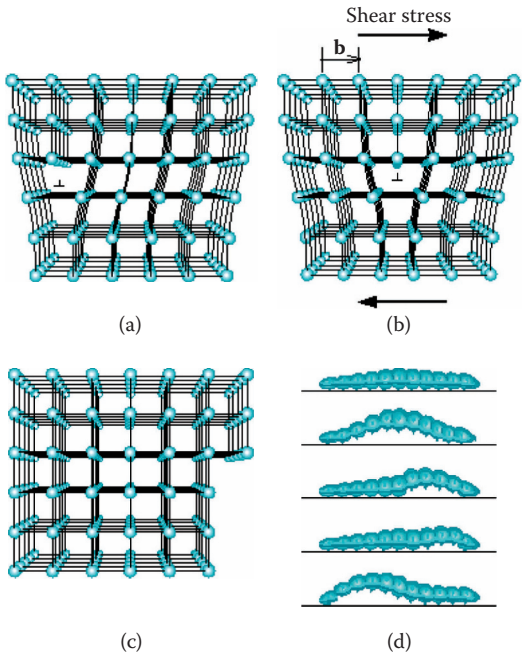


FIGURE 2.41 Dislocation movement: When a shear stress is applied to the dislocation in (a), the atoms are displaced, causing the dislocation to move one Burgers vector **b** in the slip direction (b). Continued movement of the dislocation eventually creates a step (c), and the crystal is deformed. (d) The caterpillar does not move its complete body at a single time, but it moves one segment at a time as it pulls itself forward.

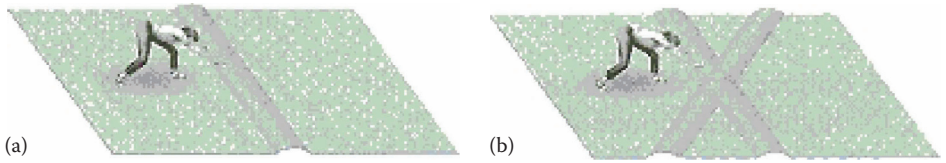


FIGURE 2.42 Moving a carpet over the floor to illustrate the effect of a line dislocation in a crystal: (a) dislocation; (b) work hardening.

a plane breaking at the same time, atoms are gliding gently in the direction of applied stress normal to dislocation lines. Thus, the direction of movement, i.e., the deformation, is the same as the Burgers vector, **b**. As a consequence **b** is also called the slip vector.

The interatomic forces in a crystal offer little resistance to the gliding motion of dislocations. An analogy is that of moving a carpet across the floor. This is difficult because of the friction developed from the contact of the whole surface of the carpet with the floor. But with a wrinkle in the carpet, as shown in Figure 2.42, the carpet can now be moved by pushing the wrinkle across the floor. The work involved is much less because only the friction between a small section of carpet and the floor has to be overcome. A similar phenomenon occurs when one plane of atoms moves past another by means of a dislocation defect.

Figure 2.43 shows how an applied shear stress, τ, exerts a force on a dislocation and is resisted by a frictional force, **F**, per unit length. The work done by the shear stress (W_τ) equals the work done by the frictional force (W_F), or:

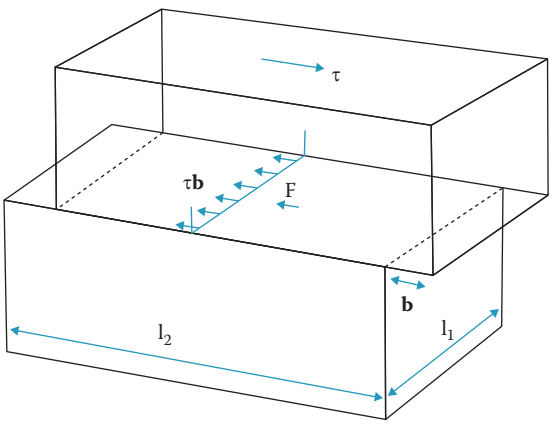


FIGURE 2.43 An applied shear stress, τ, exerts a force on a dislocation and is resisted by a frictional force, **F**, per unit length. The slip vector or Burgers vector is **b**. (Drawing by Mr. Chengwu Deng.)

$$W_\tau = (\tau\, l_1 l_2) \times \mathbf{b}$$

and (2.62)

$$W_F = (F\, l_1) \times l_2$$

With $W_\tau = W_F$, the lattice friction stress for dislocation motion is:

$$\tau = \frac{\mathbf{F}}{\mathbf{b}} \tag{2.63}$$

where **F** is the force per unit length of the dislocation and **b** is its Burgers vector or slip vector. Thus, the applied stress produces a force per unit length everywhere along the dislocation line equal to $\tau\mathbf{b}$ and perpendicular to the line element. It can be shown that:

$$\tau = \frac{\mathbf{F}}{\mathbf{b}} = \mu e^{\frac{2\pi \mathbf{a}}{\mathbf{b}}} \tag{2.64}$$

This lattice friction stress is much less than the theoretical shear strength. Dislocation motion, from Equation 2.64, is most likely on closed packed planes (large **a**, interplanar spacing) in closed packed directions (small **b**, in-plane atomic spacing). It is thus easiest to create dislocations in the closest packed crystals, and they are typically very soft and ductile.

When dislocations move it is said that slip occurred, and the lattice planes that slipped are called—imaginatively—slip planes as elucidated in Figure 2.44 for an edge dislocation and a screw dislocation. As mentioned above, the preferred slip planes are those with the greatest interplanar distance, e.g., (111) in FCC crystals, and the slip directions are those with the lowest resistance, i.e., the closest packed direction. Slip lines are the intersection of a slip plane with a free surface.

The motion of dislocations can be blocked by another dislocation, a grain boundary, or a point

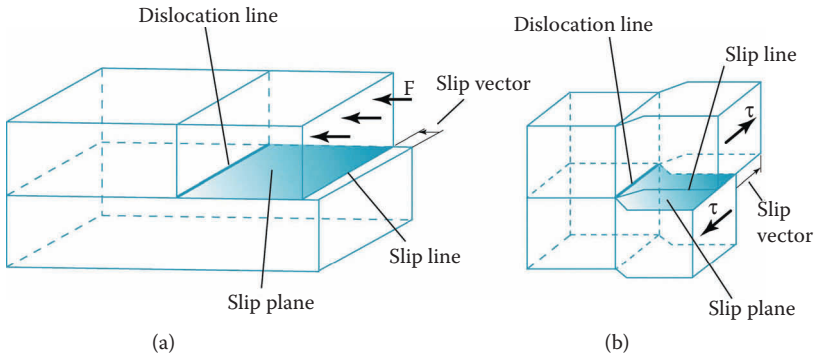


FIGURE 2.44 Schematic of slip line, slip plane, and slip (Burgers) vector for (a) an edge dislocation and (b) for a screw dislocation. (Drawing by Mr. Chengwu Deng.)

defect. Two dislocations may repel or attract each other, depending on their directions. Atoms near the core of a dislocation have a higher energy because of distortion. To minimize this energy, dislocations tend to shorten as if the line had a line tension, *T*, i.e., strain energy per unit length. The line tension is given as:

$$T \approx \frac{1}{2}\mu \mathbf{b}^2 \tag{2.65}$$

Dislocations might get pinned by interstitials and bow with a radius *R* when subjected to a shear stress like the string shown in Figure 2.45:

$$\tau \mathbf{b} L = 2T \sin\frac{\theta}{2} \tag{2.66}$$

With Equation 2.65 this leads to:

$$R = \frac{\mu \mathbf{b}}{2\tau} \tag{2.67}$$

Plastic deformation (or yielding) occurs by sliding (or slip) of parallel lattice planes past each other. Pure metals have low resistance to dislocation motion, thus they exhibit low yield strength. Adding impurities in *solution strengthening* may increase the yield strength. A well-known example is alloying Zn and Cu to form brass with a strength increase of up to 10 times over pure Cu. The bigger Zn atoms make the slip plane "rougher," thus increasing the resistance to motion. In general, impurities diffuse to dislocations and form "clouds" that pin dislocations, increasing the elastic limit. Small particles, precipitates, can also promote strengthening by impeding dislocation motion. Precipitates cause bowing of dislocations as illustrated in Figure 2.45. A critical condition is reached when the dislocation takes on a semicircular configuration between two particles separated by a distance *L*. Beyond this point a dislocation may loop and escape between the finely dispersed particles in the metal. In the semicircular situation $\tau \mathbf{b} L = 2T$ or $\tau = 2T/\mathbf{b}L$, and with Equation 2.67 $\tau = \mu \mathbf{b}/L$; thus, making *L* smaller will disadvantage dislocation looping. Finally, metals can be hardened by work hardening. Dislocations move when metals are subjected to "cold work," and their density can increase up to 10^{12} dislocations/cm^2 because of the formation of new dislocations and dislocation multiplication (see below). The consequent increasing overlap between the strain fields of adjacent dislocations gradually increases the resistance to further dislocation motion. This causes a hardening of the metal as the deformation progresses (Figure 2.42b). This effect is known as strain hardening. The effect of strain hardening can be removed by appropriate heat treatment (annealing), which promotes the recovery and subsequent recrystallization of the material. Annealing decreases the dislocation density to around 10^6 dislocations/cm^2.

Some dislocations form during the process of crystallization, but more are created during plastic deformation. Frank and Read elucidated one possible mechanism by which dislocations multiply;

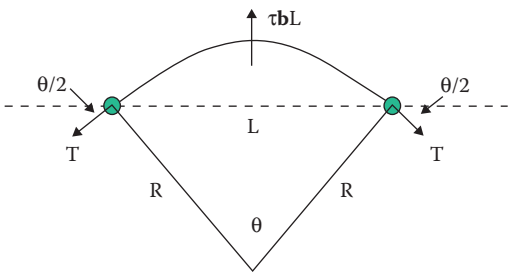


FIGURE 2.45 A pinned dislocation bows under a shear stress.

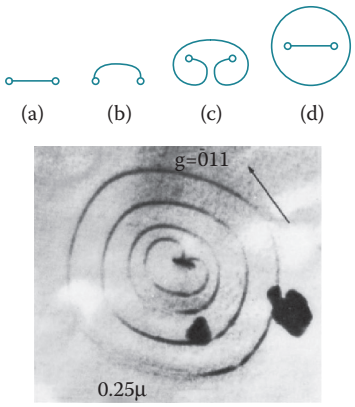


FIGURE 2.46 A Frank-Read source can generate dislocations. (a) A dislocation is pinned at its ends by lattice defects. (b) As the dislocation continues to move, the dislocation bows, eventually bending back on itself. (c) Finally the dislocation loop forms, and (d) a new dislocation is created. (e) Electron micrograph of a Frank-Read source (×330,000).

they found that a pinned dislocation (as shown in Figure 2.45), under an applied stress, produces additional dislocations. This mechanism is at least partly responsible for strain hardening. The Frank-Read dislocation multiplication mechanism is illustrated in Figure 2.46. First, a dislocation is pinned at its ends by lattice defects, and as the dislocation continues to move, it bows and eventually bends back on itself. Then the dislocation loop forms, and a new dislocation is created. Figure 2.46 also shows an electron micrograph of a Frank-Read source.

Schmid's Law, illustrated in Figure 2.47, gives the relationship between shear stress, the applied stress, and the orientation of the slip system. Slip on a given

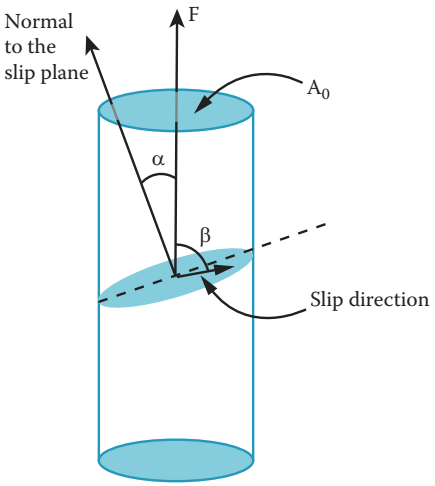


FIGURE 2.47 Geometry of slip plane, slip direction, and tensile force F. (Drawing by Mr. Chengwu Deng.)

slip system begins when the shear stress resolved on that system reaches a critical value. Consider a cylindrical crystal of cross-section A_0 under the influence of a tensile force F. Let the normal to the active slip plane make an angle α with F, and let the angle between the slip direction and F be β. The resolved shearing force, i.e., the force acting per unit area of the slip plane in the slip direction, is then given by:

$$\tau = \frac{F}{A}\cos\alpha\cos\beta \tag{2.68}$$

Because the area of the slip plane is $A/\cos\alpha$, the tensile stress per unit area normal to the slip plane is:

$$\sigma = \frac{F}{A}\cos^2\alpha \tag{2.69}$$

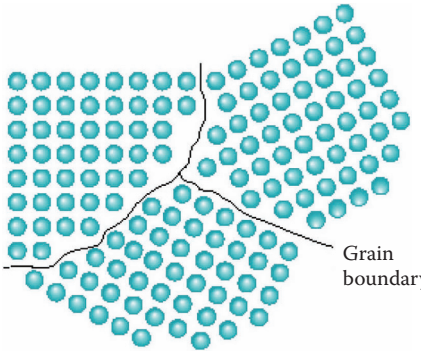
When increasing the force F, the rate of plastic flow increases very rapidly when the resolved shear stress τ reaches a critical value τ_c. In general τ_c decreases with increasing temperature and increases as a result of alloying or cold working (see above).

Besides point and line defects there are also surface defects to reckon with in single crystals. Surface defects include grain boundaries, phase boundaries, and free surfaces. The crystal structure is disturbed at grain boundaries, and different crystal orientations are present in different grains (Figure 2.48).

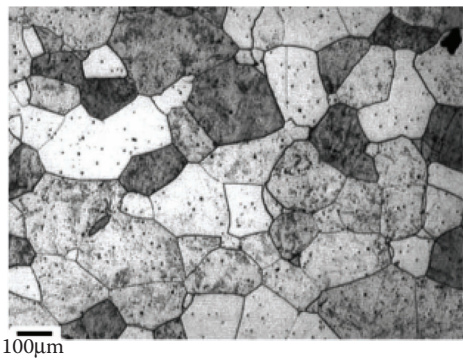
Dislocations are blocked by grain boundaries, so slip is blocked. The smaller the grain size, the larger the surface of the grain boundaries and the larger the elastic limit. The latter is expressed in the empirical Hall-Petch equation for the maximum elastic yield strength (stress at which the material permanently deforms) of a polycrystalline material:

$$\sigma_Y = \sigma_0 + \frac{K}{\sqrt{d}} \tag{2.70}$$

where d is the average diameter of the grains in micrometers, with σ_0 the frictionless stress (N/m^2) that opposes dislocation, and K a constant. The strength of a material thus depends on grain size. In a small grain, a dislocation gets to the boundary and stops, i.e., slip stops. In a large grain, the dislocation can travel farther. So small grain size equates to more strength. For example, the elastic limit of copper doubles when the grain size falls from 100 μm



(a)



(b)

FIGURE 2.48 Crystal structure is disturbed at grain boundaries. Schematic representation of grain boundaries (a) and microscope picture (b). (From Askeland D. R., and P. P. Phule, *The science and engineering of materials*, Brooks/Cole, Pacific Grove, CA, 2003.)

to 25 μm. Instead of yield strength one might also plot the hardness (H) of the material as a function of grain size, and a similar relationship is obtained (see Volume II, Figure 7.54): smaller grain size corresponds to a harder material. Indentation (hardness) testing is very common for bulk materials where the direct relationship between bulk hardness and yield strength is well known. As we will learn in Volume II, Chapter 7, in the section on thin film properties, the Hall-Petch relationship has been well established for grain sizes in the millimeter through submicrometer regimes but is less well-known in the nano regime. Based on Equation 2.70 one expects that nano-sized grains would produce materials with yet greater mechanical integrity, but in reality this is not the case! There is a reverse Hall-Petch effect, i.e., the strength of materials, from a small grain size on, starts to decrease with decreasing grain size. Plastic deformation occurs at lower and lower stresses as the grains shrink. In other words, an optimal grain size (d_c) exists as suggested by the plot in Volume II, Figure 7.54. The classic Hall-Petch relationship is based on the idea that grain boundaries act as obstacles to dislocation movement, and because dislocations are carriers of plastic deformation, this manifests itself macroscopically as an increase in material strength. The Hall-Petch behavior breaks down when the smallest dislocation loop no longer fits inside a grain. Lattice dislocation cannot be the way very small-grained materials deform. The deformation mechanism for materials with very small grains (<20 nm) is indeed different, and it has been suggested that plastic deformation in this case is no longer dominated by dislocation motion; it is believed that individual atoms migrate, diffuse, and slide along grain boundaries and through triple junctions (Y-shaped grain boundary intersections).[1]

In Volume III, Chapter 7, on scaling, we learn that the surface-to-volume ratio (S/V) of particles scales as $1/r$, where r is the characteristic dimension of the particles. The smaller a particle, the more of its atoms find themselves at its surface. A bulk solid material will typically have less than 1% of its atoms on its surface but 10-nm particles have about 15% of surface atoms (Figure 2.49). The high S/V ratio of nanoparticles makes them

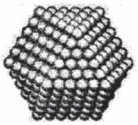
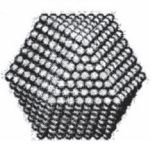
Full-Shell Clusters		Total Number of Atoms	Surface Atoms (%)
1 Shell		13	92
2 Shells		55	76
3 Shells		147	63
4 Shells		309	52
5 Shells		561	45
7 Shells		1415	35

FIGURE 2.49 Nanoparticles, clusters of atoms in shells.

more reactive as catalysts in chemical reactions and lowers their melting temperature, T_m. There is an inverse linear relationship between melting temperature and the surface-to-volume ratio. This makes sense because atoms on a surface are more easily accessed and rearranged than atoms in the bulk. As a consequence, the melting temperature of particles is always lower than the bulk. Grains in a solid are analogous to particles, and the melting temperature of solids decreases with grain size. Although the temperature dependence of σ_0 and K in Equation 2.70 can be neglected for conventional grain sizes, for grains smaller than 20 nm this assumption breaks down. In nanomaterials the melting temperature decreases because of the smaller grain size.

If we could grow a crystal without dislocations, it should approach the theoretical shear stress τ calculated in Equation 2.61. Sometimes fine metallic whiskers can be grown virtually free of dislocations, and these are very strong indeed.

In Figure 2.50 we show a picture of a zinc whisker. Metal whiskers are a crystalline metallurgical phenomenon whereby metal grows tiny, filiform hairs. The effect is primarily seen with elemental metals but is also observed with alloys. The mechanism behind metal whisker growth is not well understood but seems to be encouraged by compressive mechanical stresses, including residual stresses caused by electroplating, mechanically induced stresses, stresses induced by diffusion of different metals, and thermally induced stresses. Metal whiskers differ from metallic dendrites in several respects; dendrites are fern-shaped and grow across the surface of the metal, whereas metal whiskers are hair-like and project at a right angle to the surface. Although the precise mechanism for whisker formation remains unknown, it is known that whisker formation does not require either dissolution of the metal or the presence of an electromagnetic field. Whiskers, like carbon nanotubes, approach ideal lattices.

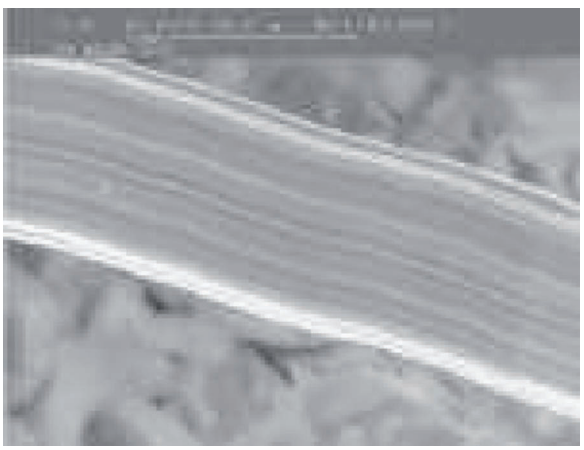


FIGURE 2.50 SEM of a zinc whisker; diameter is 10 μm.

As we are moving to smaller and smaller functional devices in MEMS and NEMS, surfaces and crystal imperfections contribute more and more importantly to overall performance and behavior. Many of the observations described here will have to be revisited when exploring nanomaterials and nanodevices. These effects will be treated in Volume II, Chapter 7 on thin films and in Volume III, Chapter 3 on nanotechnology.

Acknowledgments

Special thanks to Xavier Casadevall I Solvas, Robin Gorkin, Chengwu Deng, Kartikeya Malladi, Leyla Esfandiari, Fatima Alim, and Drs. Sean Parkin, Han Xu, and Guangyao Jia.

Appendix 2A: Plane Wave Equations

We define a plane wave,* $e^{i\mathbf{k}.\mathbf{r}}$, as a wave that is constant in a plane perpendicular to $\mathbf{k}$ (in rad/m) and is periodically parallel to it, with a wavelength $\lambda = 2\pi/\mathbf{k}$, where $\mathbf{k}$ is the wave number (because it measures the number of wavelengths in a complete cycle) with a value of $2\pi/\lambda$ and where $\omega = 2\pi\nu = c\mathbf{k}$ (free space) is the period of the wave (see also Table 2A.1). A bit simpler: a plane wave is a continuous wave (CW) whose amplitude and phase are constant in the directions transverse to the propagation direction.

* The complex exponential representation for periodic functions is convenient for adding waves, taking derivatives of wave functions, and so on. It is equivalent to a linear combination of a sine and cosine function because $e^{i\theta} = \cos\theta + i\sin\theta$.

Table 2A.1 Plane Wave Parameters

Parameter	Symbol/Value
Amplitude	$\mathbf{E}_m, \mathbf{B}_m$
Phase	$\phi = kx - \omega t$
Velocity	$v = \frac{\omega}{\mathbf{k}}$
Wavelength	λ
Period	T
Wave vector	$\mathbf{k} = \frac{2\pi}{\lambda}$
Angular frequency	$\omega = \frac{2\pi}{T}$
Wave number	$\bar{\mathbf{k}} = \frac{1}{\lambda}$
Cyclic frequency	$v = \frac{1}{T}$

The amplitude of a wave propagating in the x-direction is mathematically introduced as

$$\psi = \psi_0 \sin(kx + \omega t + \delta)$$
$$\omega = 2\pi f, \text{ the cyclic frequency}$$
$$k = \frac{2\pi}{\lambda}, \text{ the wave vector}$$
$$\delta, \text{ the phase at } t = 0 \text{ and } x = 0$$

Calculations are greatly simplified by using complex numbers:

$$\psi = \psi_0 \text{Im } e^{i(kx+\omega t+\delta)}$$
$$i = \sqrt{-1}$$
$$|e^{ix}| = 1$$
$$\text{Im } (e^{ix}) = \sin x$$
$$\text{Re } (e^{ix}) = \cos x$$
$$\sin x = \frac{e^{ix} + e^{-ix}}{2i}$$
$$\cos x = \frac{e^{ix} + e^{-ix}}{2}$$

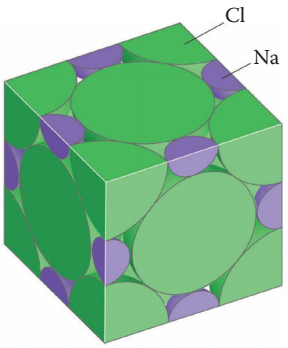
Questions

2.1: Scientists are considering using nanoparticles of magnetic materials such as iron-platinum (Fe-Pt) as a medium for ultrahigh density data storage. Arrays of such particles potentially can lead to storage of trillions of bits of data per square inch—a capacity that is 10 to 100 times higher than any current storage devices such as computer hard disks. If these scientists consider iron (Fe) particles that are 3 nm in diameter, what is the number of atoms in one such particle?

2.2: Assuming that silica (SiO_2) has 100% covalent bonding, describe how oxygen and silicon atoms in silica (SiO_2) are joined.

2.3: What is the difference between lattice and basis and between unit cell and primitive cell?

2.4: What are the net numbers of Na^+ and Cl^- ions in the NaCl unit cell represented below? The crystal is an example of which type of cubic lattice? Identify the atom positions of the Na and Cl atoms in the NaCl structure.



2.5: Consider the plane defined by the three points, P1(2,4,–3), P2(–1,2,1), and P3(3,0,–2). Calculate the points where this plane intersects with the axes and derive the Miller indices associated with this plane.

2.6: Calculate (a) the angle between [111] and the direction normal to (111) plane in a simple cubic crystal and (b) the angle between the [121] direction and the direction normal to (113) plane in a simple cubic crystal.

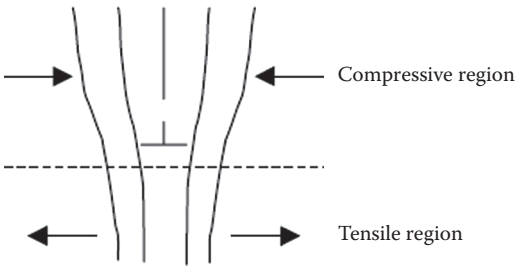
2.7: A signal $x(t)$ has a Fourier transform $X(f)$. Express the Fourier transforms of x_1 in terms of $X(f)$, the function $x_1(t) = x(3 - t)$.

2.8: In an x-ray set-up with a crystal-to-detector distance of 100 mm (D), you find that the highest resolution reflection is at $x = 20$ mm, $y = 15$ mm, relative to the direct beam position. The wavelength $\lambda = 1.54$ Å. What is the Bragg angle of this reflection? What is the d-spacing of the crystal? If the detector is

circular and has a radius of 100 mm, and you would like to collect data so that the highest resolution reflection is at the detector edge, would you move the detector closer to or further away from the crystal?

2.9: Describe the difference between a reciprocal lattice and a real one.

2.10: Edge dislocations may be used to getter impurities in semiconductors. In an edge dislocation, there are always two regions, a compressive region, where the layer is inserted, and a tensile region (see figure below). Which of these two regions will best accommodate (via elastic interactions) *substitutional* atoms whose radius is larger than that of the host atoms?

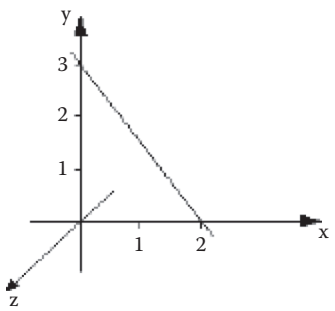


2.11: Show that the reciprocal lattice of a face-centered cubic (FCC) lattice is a body-centered cubic (BCC) lattice or, conversely, that the reciprocal lattice of a BCC lattice is an FCC lattice.

2.12: An electron moves with speed u = 0.7c. Calculate its total energy and its kinetic energy in eV.

2.13: Calculate the number of atoms in a 100 μm long Ag line (1 μm wide and 1 μm high). If using STM we put one atom down per second, how long will it take to finish this Ag line?

2.14: What is the Miller index for the plane shown below?



2.15: What is the number of nearest neighbors for the following crystal lattices:
(a) simple cubic (SC)
(b) face-centered cubic (FCC)
(c) body-centered cubic (BCC)

2.16: Calculate the angle θ of reflection for an x-ray experiment with λ = 1.54 Å, for a cubic crystal with a lattice parameter of **a** = 5 Å.

2.17: X-rays with wavelength 1.54 Å are "reflected" from the (110) planes of a cubic crystal with unit cell **a** = 6 Å. Calculate the Bragg angle, θ, for all orders of reflection, *n*.

2.18: What is the closest packed crystal? Simple cubic (SC), body-centered cubic (BCC), or face-centered cubic (FCC)?

2.19: Is five-fold symmetry ever found in crystal lattices? Why or why not?

2.20: Calculate the number of atoms in 100 g of silver.

2.21: The derivation of Bragg's law results in $n\lambda = 2d\sin\theta$. What does *n* represent and why is it usually omitted? Can you give an example to show why n is not needed?

2.22: Calculate the packing factor in a FCC lattice.

2.23: What is the rule for determining the *slip direction* in a close-packed material?

2.24: What causes work hardening?

2.25: How can we determine the direction cosine between two vectors?

2.26: How can we identify the direction of slip in a crystal (Burgers vector)?

2.27: Why do very thin metal wires/whiskers exhibit very high strengths?

2.28: Why does the strength of glass fibers increase as the diameter goes down?

2.29: Find the angle between the planes 110 and 100 in a simple cubic crystal.

2.30: Could you use x-ray diffraction to determine the coefficient of thermal expansion (i.e., change in length of a material due to change in temperature)?

2.31: Diamond and graphite are examples of which type of crystalline solids: molecular, covalent network, ionic, or metallic?

2.32: What value does the atomic scattering factor f approach as 2θ approaches 0?

2.33: How does light affect the color of a crystal?

Further Reading

Vectors and Tensors

Danielson D. A. 1997. *Vectors and tensors in engineering and physics.* New York: Addison-Wesley Publishing Company, Inc.

Crystallography

Books reviewed or listed in *Crystallography News*. See http://img.cryst.bbk.ac.uk/bca/cnews/books/index.html.

The International Union of Crystallography (IUcr) produces a number of books on crystallography, in association with Oxford University Press and Springer. See http://www.iucr.org/iucr-top/publ/index.html.

Hannay, N. B. 1967. *Solid-state chemistry.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Ladd, M. F. C., and R. A. Palmer. 2003. *Structure determination by X-ray crystallography.* 4th ed. New York: Springer.

Sands, D. S. 1969: *Introduction to crystallography.* New York: W.A. Benjamin, Inc.

Solid-State Physics

Brown, F. C. 1967. *The physics of solids.* New York: W.A. Benjamin, Inc.

Dekker, A. J. 1970. *Solid state physics.* New York: Prentice-Hall, Inc.

Kittel, C. 2005. *Introduction to solid state physics*. 8th ed. New York: John Wiley & Sons, Inc.

"Phase Problem"

Eisenberg, D., and D. Crothers. 1979. *Physical chemistry: with applications to the life sciences*. San Francisco: Benjamin-Cummings/Addison Wesley Longman.

Glusker, J. P., and K. N. Trueblood. 1972. *Crystal structure analysis: a primer.* New York: Oxford University Press.

Holmes K. C., and D. M. Blow. 1965. *Use of X-ray diffraction in the study of protein and nucleic acid structure*. New York: Krieger Pub Co/Interscience.

Wilson, H. R. 1966. *Diffraction of X-rays by proteins, nucleic acids, and viruses.* New York: St. Martin's Press.

Dislocations

Hull, D. 1965. *Introduction to dislocations*. Oxford, UK: Pergamon Press.

Weertman, J., and J. R. Weertman. 1964. *Elementary dislocation theory.* New York: MacMillan.

Reference

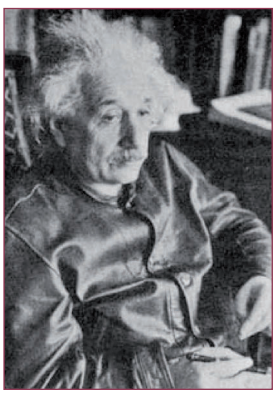
1. Schiotz, J., and K. W. Jacobson. 2003. A maximum in the strength of nanocrystalline copper. *Science* 301:1357–59.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

3

Quantum Mechanics and the Band Theory of Solids

Some of the actors in this chapter: (a) Fermi, (b) Schrödinger, (c) de Broglie, (d) Feynman, (e) Born, (f) Bohr, (g) Rutherford, (h) Dirac; and below is Einstein.



Outline

The only reason for time is so that everything doesn't happen at once.

Albert Einstein

Nothing is real.

John Lennon, 1940–1980

Introduction

As we came to understand from Chapter 2, when atoms are brought together in crystals, their outermost electrons are influenced by a periodic potential. Therefore, the possible electron energies form bands of allowed values separated by bands of forbidden values. This band structure is of fundamental importance in explaining the properties of metals, semiconductors, and insulators. The most accepted and most accurate theory in modern physics today, to predict physical, mechanical, chemical, and electrical properties of atoms, molecules, and solids—including the band theory of solids—is quantum mechanics.

At the end of the nineteenth century, physicists thought they had a good grasp of the physical world with mechanics (e.g., Newton, Hamilton, and Lagrange), statistical mechanics (mostly Boltzmann and Gibbs), hydrodynamics (Stokes), and electrodynamics (Maxwell). There had been a growing unease, though, about the incapability of classical theories to explain a wide variety of experiments, such as the magnitude of the electrical and thermal conductivity of metals and the heat capacity of metals and insulators. Confidence was further challenged by a series of new discoveries: radioactivity (1896), the electron (1897), the quantum (1900), the photoelectric effect (1887), and x-rays (1895). The result was the development of a set of new theories, all explaining extreme aspects of nature better than classical theories. Although the classical theories worked well at everyday velocities and scales, at the extremes, new theories were needed. For very fast phenomena, special relativity worked better; for very small particles, quantum mechanics; and for very large phenomena, general relativity was demonstrated to be superior.

Before introducing quantum mechanics, we start out by detailing the areas where classical physics started faltering at the end of the nineteenth century. We specifically describe the shortfalls of the classical theories in explaining electrical and thermal conductivity of metals, heat capacity, the observation of a positive Hall effect, the ultraviolet catastrophe, and the photoelectric effect. All the above phenomena follow logically from quantum theory.

The relevance of quantum mechanics in the context of ICs and NEMS cannot be underestimated, and the profound implications of quantum physics for nanoelectronics and NEMS are a recurring topic throughout this book. In the IC and NEMS world, we are moving fast into the realm of quantum mechanics. Moore's law might remain valid until about 2020, but by then the scale of electronic components will be at the molecular/atomic level, and hence can no longer be described by classical mechanics. Quantum computing and nanotechnology, including nanotubes, nanowires, biological nanostructures, and quantum dots, all require some grounding in quantum mechanics to be understood at all. Quantum mechanics must now become a familiar tool not only to physicists but also to materials scientists, biologists, and electrical, mechanical, and bioengineers.

Classical Theory Starts Faltering

Introduction

In this section, we cover examples where microscopic classical models started failing at explaining a wide variety of experimental results obtained in the late nineteenth century. The examples include early models devised to explain electrical conductivity in metals, heat capacity of metals and insulators, the temperature dependence of electrical conductivity and heat capacity, thermal conductivity, the Hall effect, blackbody radiation, and the photoelectric effect. The failing of each model ultimately originated in the mistaken assumption that lattice vibrations (phonons), electrons, and photons could all take on continuous energy values, an error that quantum mechanics corrects.

Electronic Conductivity

DC Electrical Conductivity

Paul Drude (1863–1906) (Figure 3.1) was intrigued with the huge resistivity range in materials—from $10^{18}\ \Omega\cdot$cm for fused quartz to $10^{-6}\ \Omega\cdot$cm for silver (Figure 3.2)—and developed one of the first models to explain electrical conductivity in metals.

Modern condensed matter physics really started with the discovery of the electron by J.J. Thompson

FIGURE 3.1 Paul Drude (1863–1906).

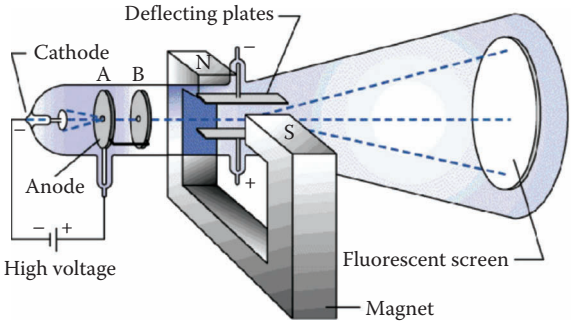


FIGURE 3.3 J.J. Thomson demonstrated that an electron behaves as a particle using an apparatus as shown here (A and B are anodes). The electron stream is deflected by electric and magnetic fields.

in 1897. In a series of experiments performed before the year 1900, Thompson demonstrated that electrons behave as particles of mass m that carry a fixed amount of negative electrical charge, e. These particles move in trajectories governed by the laws of electricity, magnetism, and classical mechanics, and Thompson determined the ratio e/m directly using an instrument as shown in Figure 3.3. Thompson received the Nobel Prize for his work on the electron in 1906. It was in the famous oil-drop experiment of Robert Millikan (1868–1953), carried out in 1909, that the size of the charge on an electron was finally established. Millikan also determined that there was a smallest "unit" charge, or that charge is "quantized."

Drude relied on the newly introduced concept of electrons to postulate a theory of metallic conductivity. His work was before the development of quantum mechanics, so he relied on classical physics models only. He introduced the idea of a free electron gas (FEG), in which he assumed that an electron gas surrounds the positive ion cores of the metal and that electron-ion interactions are negligible. Applying the kinetic theory of gases to the free electron gas, he had electrons only moving in straight lines and colliding only with ion cores, thus neglecting electron-electron interactions. Drude also envisioned collisions in which electrons instantaneously lose all the energy they previously gained from an electric field. He approximated the mean free path of the electrons, λ, with the interionic core spacing in the solid lattice, a. It should be noted that in reality electron densities (10^{22}–10^{23} cm^{-3}) are thousands of times greater

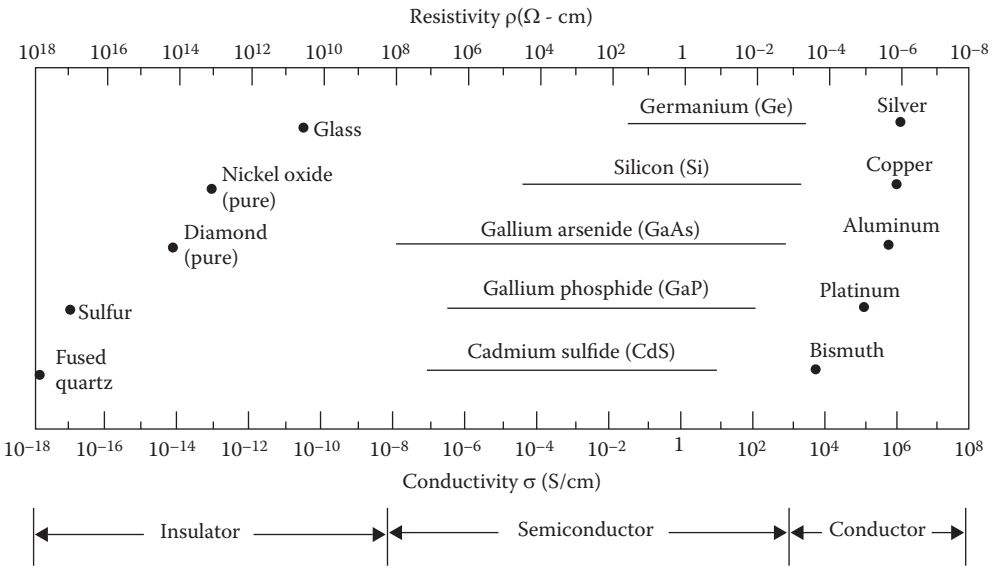


FIGURE 3.2 Conductivity/resistivity range for common materials.

than those of a gas at normal conditions, and the assumption that there are no strong electron-electron and electron-ion interactions is wrong. In spite of this, the model successfully explains the form of Ohm's law. However, it does fail to explain many other important aspects of conductors, such as the exact magnitude of electrical conductivity and heat capacity and their temperature dependence, or the relationship between electrical and thermal conduction.

In what follows we apply macroscopic boundary conditions to a piece of metal. In the nanoworld, as we will learn, things are different: with a macroscale conductor wire, the mean free path of the electrons, λ, is very small compared with the wire length, and the motion of electrons is diffusive. But, with a nanowire, the wire is short compared with λ and the electron motion may turn ballistic.

We start with the experimental observation that the current in a conductor is proportional to the applied voltage ($V \propto \mathbf{I}$), i.e., Ohm's law. If an electrical field **E** (V/m) is applied over a conductor of length *L*, a charge, *Q*, flows and an electrical current, $\mathbf{I} = dQ/dt$, results [SI unit: 1 ampere (A) = 1 coulomb per second (C/s)]. The current density is given as $\mathbf{J} = \mathbf{I}/A$ or $d\mathbf{I}/dA$, with *A* the cross-sectional area of the current pathway or also:

$$\mathbf{I} = \int \mathbf{J}\, dA \tag{3.1}$$

where **J** is given by (see Chapter 2):

$$\mathbf{J} = \sigma\mathbf{E} \tag{2.2}$$

or also (see Chapter 2):

$$\mathbf{E} = \rho\mathbf{J} \tag{2.4}$$

σ is the electrical conductivity, and ρ is the electrical resistivity ($\sigma = 1/\rho$). The SI unit for resistivity is in ohm·meter (Ω·m) and ohm = volt/ampere (V/A); the SI unit for conductivity is in siemens per meter (S/m^{-1}) and siemens = ampere/volt = ohm^{-1}. The conductivity, σ, and resistivity, ρ, are intrinsic characteristics of a material and are independent of sample geometry but are linked to the crystal structure as we may glean from Equations 2.3–2.7 in Chapter 2. From Equation 2.2, σ measures the current density for a given electric field. For now, we assume

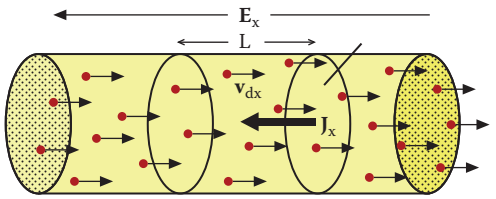


FIGURE 3.4 Electron conduction in a section of metal wire in the presence of an electrical field. The average distance traveled by an electron is the mean free path λ, which Drude assumed to be equal to the lattice constant, *a*.

an isotropic solid, and because $\mathbf{J} = \mathbf{I}/A$ and $\mathbf{E} = V/L$, Equation 2.2 can be rewritten as $\mathbf{J} = \mathbf{I}/A = \sigma\, V/L$ or:

$$V = \mathbf{I}(L/\sigma A) = \mathbf{I}(\rho L/A) = \mathbf{I}R \tag{3.2}$$

i.e., the familiar Ohm's law, with the resistance, *R*, linked to the geometry of the sample (*L* and *A*). The SI unit for resistance is ohm (Ω).

Drude's microscopic interpretation of Ohm's law can be understood from an inspection of Figure 3.4, where we consider the conduction of electrons in a metal wire on application of an electrical field in the *x*-direction, $\mathbf{E}_x$. There are *n* electrons per unit volume (electron density), and they all move in the direction of the current, *I*, with a drift velocity in the *x*-direction, $\mathbf{v}_{dx}$. The drift velocity, $\mathbf{v}_{dx}$, is the net motion of electrons opposite to the electrical field (Figure 3.5). The number of electrons crossing area *A* in a time *dt* is $nA\mathbf{v}_{dx}dt$. In the time segment *dt*, the electrons have traveled a distance *L* along the wire. The current density in the *x*-direction, $\mathbf{J}_x$, is then $\frac{\mathbf{I}}{A}$ or $\frac{\Delta Q}{A\Delta t} = \frac{(ne)(AL)}{AL/\mathbf{v}_{dx}}$, and with *Q* the charge we calculate:

$$\mathbf{J}_x = en\mathbf{v}_{dx} \tag{3.3}$$

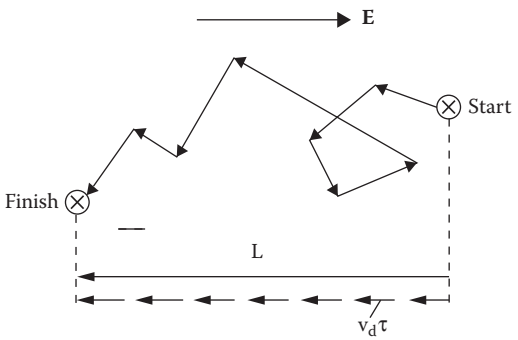


FIGURE 3.5 Drift of an electron opposite the electrical field. During an average time τ, the electron travels a mean free path $\lambda = v_{dx}\tau$. In a time *dt*, the electrons have traveled a distance *L* along the wire.

where e is the charge of the electron. By convention, a positive charge moves with the electrical field, and Equation 3.3, more correctly, should read $\mathbf{J}x = env_{dx}$, with $\mathbf{J}_x$ and $\mathbf{v}_{dx}$ vectors with a direction and magnitude. For simplicity, we often leave the vector notation out, and introduce it only when we want to draw special attention to it. Drude treated the free electron gas as a gas of molecules where distribution of velocities follows the Maxwell-Boltzmann distribution (see below). In such a gas, at T = 0, all the free electrons in a conductor have zero kinetic energy. When heating the conductor, the lattice ions acquire an average kinetic energy of $\frac{3}{2} k_B T$, where k_B is the Boltzmann constant. This average energy is imparted to the electron gas by collisions between electrons and lattice ions. The latter is a consequence of the equipartition theorem. At ordinary temperatures (~300 K) the mean kinetic energy of the electrons based on this model is about 0.04 eV. From this mean kinetic energy, we approximate the thermal velocity, v_{th}, of the electrons in a metal from the root mean square (rms speed) or $v_{rms} = \sqrt{\frac{3k_B T}{m_e}}$ (see Equation 3.20 below), which is slightly larger than the average v_{avg} or mean speed $\bar{v} = \sqrt{\frac{8k_B T}{\pi m_e}}$ (see Equation 3.21 below), as $v_{th} = \bar{v} \approx v_{rms} = 1.57 \times 10^5$ m/s or almost 1000 km/s (this is about 1% of the speed of light!). However, when calculating $\mathbf{v}_{dx}$ in Equation 3.3 for a current of 1 A, one obtains velocities of the order of 10^{-4} m/s or only 0.1 mm/s! The reason for the huge difference between drift and thermal velocity is that electrons travel at fast thermal velocities for a short, average time, τ, and then "scatter" because of collisions with atoms, grain boundaries, impurities, or material surfaces (especially in very thin films or small particles). The drift velocity is also not how fast "electricity travels," for electric fields travel essentially at the speed of light. To reconcile this discrepancy, think of electrons in a wire as a pipe full of water; when a little water enters one end of the pipe, almost immediately some water flows out at the other end.

The very small drift velocity caused by the electric field has essentially no effect on the very large mean speed of the electrons: in other words, v_{th} does not depend on **E**. Drude assumed that all of the electrons' forward velocity is reduced to zero after each collision and must then be accelerated again by the electrical field.

The forces exerted on a colliding electron in an electric field, **E**, are given by Newton's second law, according to which forces give rise to a change in the momentum of particles [$(\mathbf{F} = d(m_e \mathbf{v}_{dx})/dt = d\mathbf{p}_x/dt$] or:

$$\frac{d\mathbf{p}_x}{dt} = -e\mathbf{E}_x + \mathbf{F}_{coll} = -e\mathbf{E}_x - \frac{\mathbf{p}_x}{\tau}$$
$$= -e\mathbf{E}_x - \frac{m_e \mathbf{v}_{dx}}{\tau} \qquad (3.4)$$

where $-e\mathbf{E}_x$ is the force on the electron garnered from the electrical field, m_e the mass of an electron in vacuum, $\mathbf{F}_{coll}$ the collision force, and $\boldsymbol{p}_x$ the electron's momentum in the x-direction ($=m_e \mathbf{v}_{dx}$). Because

$$-e\mathbf{E}_x - m_e \mathbf{v}_{dx}/\tau = 0 \qquad (3.5)$$

the result is a constant average velocity:

$$\mathbf{v}_{dx} = -\frac{e\mathbf{E}_x \tau}{m_e} \qquad (3.6)$$

where we introduced a minus sign because, as remarked above, the drift velocity is opposite the electrical field **E**. Substituting this result in Equation 3.3 and generalizing for three directions, we obtain the famous Drude result (ignoring vector notation):

$$\sigma = \frac{ne^2\tau}{m_e} \quad \text{and} \quad \rho = \frac{m_e}{ne^2\tau} \qquad (3.7)$$

Scatter time τ is also known as the relaxation time, the collision time, or the mean free time a randomly picked electron travels before the next collision. Scatter time τ decreases with increasing temperature T, i.e., more scattering at higher temperatures leads to higher resistivity in a metal. During an average time τ, electrons travel a mean free path λ, i.e., $\lambda = v_{th}\tau$, and in terms of the mean free path and the mean speed the resistivity is given as:

$$\rho = \frac{m_e v_{th}}{ne^2\lambda} \qquad (3.8)$$

According to Ohm's law, the resistivity is independent of the field, and in Equation 3.8 only λ and v_{th} could possibly be dependent on the field. But we

just saw that the very small drift velocity caused by the electric field has essentially no effect on the very large mean speed of the electrons or v_{th} does not depend on **E**. Drude assumed the electron mean free path, λ, to be equal to the lattice constant, a, which is of the order of 1 nm, and of course also independent of the electric field **E**; neither λ nor v_{th} depends on **E** in accordance with Ohm's law. In the absence of an electric field, the electrons perform a random, thermal motion, and there is no conduction, but when an electric field is applied, electrons move into a direction opposite to the field, thus generating a current. The value of the resistance is finite because electrons collide with the lattice ions, and they are stopped frequently in their tracks before being accelerated again. This simple model does explain correctly the form of Ohm's law.

From Equation 3.7 there are two contributions to the conductivity. One is the number of charges (ne), and the second is how easily those charges can be accelerated ($e\tau/m_e$). To really understand the differences between the numbers of "free charges" in different materials requires the quantum description introduced below. However, let us accept for now that in a material such as silver, the number of electrons that are free to move equals $5.8 \times 10^{28}/m^3$ and that in zinc it is even larger, $1.3 \times 10^{29}/m^3$. From these numbers one would expect zinc to be the better conductor! However, as we shall see, this is not the end of the story. The second contribution to the conductivity in Equation 3.7 is the charge carrier's drift mobility, μ_e (in the case of electrons), which, in cm^2/Vs, is given as:

$$\mu_e = \frac{e\tau}{m_e} \tag{3.9}$$

Therefore, the conductivity is the product of the number of free charges and the mobility of those charge carriers:

$$\sigma = \mu_e ne \tag{3.10}$$

The drift mobility, μ_e, may also be defined as the drift velocity per unit applied electric field, or:

$$\mu_e = \frac{v_d}{E} \tag{3.11}$$

The drift mobility, from Equation 3.9, is linked in turn to the mean scattering time between collisions, τ. Coming back to the comparison of the conductivity of silver and zinc, typical mobilities in silver are much higher than those in zinc so that the electrical conductivity is higher, even though zinc has more electrons. This makes silver the better conductor after all. Once we have introduced quantum mechanics, it will become clear that the mean scattering time, τ, has nothing to do with the stagnant lattice ions in the crystal but is determined by lattice vibrations, imperfections, and impurities instead. Consequentially, the mobility of a metal can be reduced by introducing defects (e.g., kinking a wire) or by increasing the number of lattice vibrations, i.e., phonons (by raising the temperature).

The Maxwell-Boltzmann Distribution

The next question is, how does resistivity of a metal depend on temperature? To answer this we calculate, à la Drude, how the drift velocity, v_d, depends on temperature. With electrons behaving like an ideal gas, the distribution of electron speeds is described by a Maxwell-Boltzmann (MB) distribution. In a MB distribution, the probability of finding particles in a particular energy state varies exponentially as the negative of the energy divided by k_BT or:

$$f_{MB}(E) = Ae^{-\frac{E}{k_BT}} \tag{3.12}$$

where A is a normalization constant, and $e^{-\frac{E}{k_BT}}$ is called the Boltzmann factor. Based on Equation 3.12, at a given temperature, particles are distributed among all available levels, and the ground state always contains the bulk of the particles, whereas other levels will contain exponentially less. The Maxwell-Boltzmann distribution is illustrated in Figure 3.6.

The number of particles per unit volume, i.e., density $n(E)$, that have energies between E and $E + dE$ is $n(E)dE$, where $n(E) = G(E)f_{MB}(E)$. The function $G(E)dE$ is the total number of possible (allowed-to-occupy) energy states per unit volume with energies between E and $E + dE$. It is also referred to as the density of state function, often abbreviated as DOS. The other function, $f_{MB}(E)$, is the Maxwell-Boltzmann distribution, representing the probability function

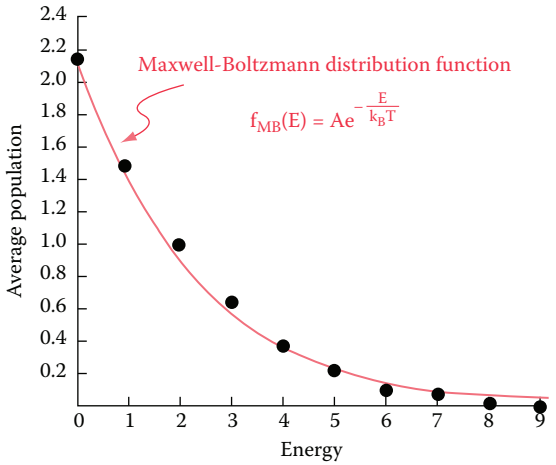


FIGURE 3.6 The Maxwell-Boltzmann distribution function: $f_{MB}(E) = Ae^{-\frac{E}{k_BT}}$.

of occupancy of a state with energy E. Because $E = mv^2/2$, we may, using Equation 3.12, rewrite $n(E)dE$ as:

$$n(E)dE = G(E)Ae^{-\frac{mv^2}{2k_BT}}dE \qquad (3.13)$$

To find the number of gas molecules with a speed in the range dv, irrespective of the direction of the velocity, we need the speed distribution function, $n(v)$. The number of allowed states (velocities) between v and $v + dv$, i.e., the number of states between a sphere of radius v and a sphere of radius $v + dv$, as illustrated in Figure 3.7, is:

$$4\pi v^2 dv = G(E)dE \qquad (3.14)$$

For every value of v there is a value of E in $G(E)$, and substituting Equation 3.14 in Equation 3.13 we obtain:

$$n(E)dE = n(v)dv = A4\pi v^2 e^{-\frac{mv^2}{2k_BT}}dv \qquad (3.15)$$

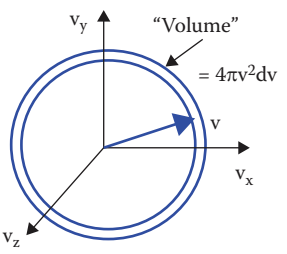


FIGURE 3.7 The velocity is a vector in three-dimensional space, and one needs to take into account that, with the increase of the magnitude of v, the space accessible to a particle increases.

To calculate the value of A, we remember that we can derive the number of molecules per unit volume or volume density of all particles (all velocities/energies) as:

$$\frac{N}{V} = \int_0^\infty n(v)dv$$

or

$$A = \frac{N}{V}\left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} \qquad (3.16)$$

$$\left(\text{since } \int_0^\infty v^2 e^{-av^2}dv = \frac{1}{4a}\sqrt{\frac{\pi}{a}}\right)$$

From Equations 3.15 and 3.16, the Maxwell-Boltzmann speed distribution is derived as:

$$n(v)dv = \frac{4\pi N}{V}\left(\frac{m}{2\pi k_BT}\right)^{\frac{3}{2}} v^2 e^{-\frac{1}{2}\frac{mv^2}{k_BT}}dv \qquad (3.17)$$

This expression gives us the speed distribution of n particles as a function of their mass and the temperature: it represents the probability that a particle has a speed in the range v to $v + dv$ as illustrated in Figure 3.8. As the temperature increases, the curve

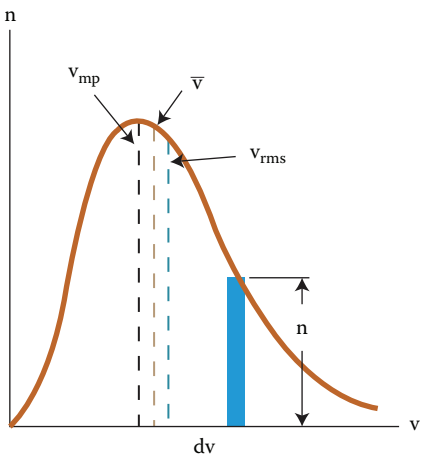


FIGURE 3.8 The Maxwell-Boltzmann distribution of speeds of particles as a function of the temperature and mass of the particles. The most probable speed v_{mp} is the speed at which the distribution curve reaches a peak, $\bar{v}$ is the average speed, and the root mean square speed is v_{rms}.

broadens and extends to higher speeds. Using $dE = mvdv$ ($E = mv^2/2$), the Maxwell-Boltzmann distribution of kinetic energies is:

$$n(E)dE = \frac{2N}{V\sqrt{\pi}}\left(\frac{1}{k_B T}\right)^{\frac{3}{2}}\sqrt{E}\, e^{-\frac{E}{k_B T}}dE \qquad (3.18)$$

Comparing Equation 3.18 with Equation 3.13 yields the 3D density of states function for gas molecules:

$$G(E) = \frac{2N}{V\sqrt{\pi}}\left(\frac{1}{kT_B}\right)^{\frac{3}{2}}\sqrt{E} \qquad (3.19)$$

This function increases smoothly with the square root of the energy.

The most probable speed $v_{mp} = \sqrt{\frac{2k_B T}{m}}$ is the speed at which the speed distribution (Equation 3.17) reaches a peak; this can be calculated from $\frac{dn(v)}{dv}$, setting it to zero, and solving for v. Although Equation 3.17 gives the distribution of speeds or, in other words, the fraction of molecules having a particular speed, we are often more interested in quantities such as the root mean square speed or the average speed of the particles rather than the actual distribution. The root mean square speed is given as:

$$v_{rms} = \sqrt{\frac{\int_0^\infty v^2 n(v)dv}{N/V}} = \sqrt{\frac{3k_B T}{m}} \qquad (3.20)$$

and the average or mean speed $\bar{v}$ is given as:

$$\bar{v} = \sqrt{\frac{\int_0^\infty vn(v)dv}{N/V}} = \sqrt{\frac{8k_B T}{\pi m}} \qquad (3.21)$$

In the case in which the Maxwell-Boltzmann distribution is applied to electrons rather than gas molecules, we must replace m with m_e, the mass of an electron. The Maxwell-Boltzmann distribution, shown in Figure 3.8, is a classical distribution of particles; in quantum statistics, other particle distribution functions are introduced. In the Maxwell-Boltzmann distribution function, it is assumed that all the particles are distinguishable; particles are physically identical but distinguishable in position and trajectory. Particles are considered distinguishable if the distance separating them is large compared with their de Broglie wavelength (see below). Another way of saying this is that the average distance between particles must be large compared with the quantum uncertainty. For an ideal gas, this criterion is certainly fulfilled, but it is not true for electrons and, as we remarked before, this is ultimately the reason why the Drude model fails to explain conductivity correctly: the Maxwell-Boltzmann distribution is not valid for a collection of electrons.

Drude Fails

Above we saw that during an average time τ, electrons travel a mean free path λ, i.e., $\lambda = v_{th}\tau$, and because Drude assumed the electron mean free path, λ, to be equal to the lattice constant, a, which is of the order of 1 nm, this yields a typical value for τ of about 10^{-14} s (because we calculated $v_{th} = \bar{v} \approx v_{rms} = 1.57 \times 10^5$ m/s). Based on Equations 3.8 and 3.21, where we replaced m with m_e and used $\bar{v}$ for calculating v_{th} (Equation 3.21), one derives for the resistivity:

$$\rho = \frac{1}{\sigma} = \frac{m_e v_{th}}{ne^2\lambda} = \frac{m_e}{ne^2 a}\sqrt{\frac{8k_B T}{\pi m_e}} \qquad (3.22)$$

Using the lattice constant, a, for the mean free path and the Maxwell-Boltzmann equation at T = 300 K to calculate v_{th}, values for the resistivity of a metal are obtained that are six times too large. In addition, from Equation 3.22 the temperature dependence of resistivity is determined by v_{th}, which in this model is proportional to $\sqrt{T}$. In practice, the temperature dependence of the resistivity is represented by the empirical relationship:

$$\rho = \rho_0 + \alpha T \qquad (3.23)$$

where ρ_0 is the resistivity at a reference temperature, usually room temperature, and α is the temperature coefficient. An expanded version of Equation 3.23 is known as the Matthiessen rule. This rule is just an approximation (a rule of thumb that works pretty well), not a true physical law. According to Matthiessen rule, the resistivity can be expressed as a sum of

TABLE 3.1 Resistivity Values of Common Metals

Material	ρ_0 ($\mu\Omega$·cm)	α ($\mu\Omega$·cm/K)*	Resistivity at 100°C ($\mu\Omega$·cm)
Aluminum	2.284	0.00390	3.0640
Copper, annealed	1.7241	0.00393	2.5101
Copper, hard-drawn	1.771	0.00382	2.5350
Brass	7.000	0.00200	7.4000
Gold	2.440	0.00340	3.1200
Iron	9.710	0.00651	11.0120
Lead	20.6480	0.00336	21.3200
Nickel	11.000	0.00600	12.2000
Silver	1.590	0.00380	2.3500
Steel	10.400	0.00500	11.4000
Nichrome	100.000	0.00040	100.0800
Platinum	10.000	0.00300	10.6000
Tungsten	5.600	0.00450	6.5000

*Determined at 25°C.

terms resulting from (nearly) independent contributions, for example, $\rho = \rho_0 + \rho_T T + \rho_I C + \rho_{other}$. Here, ρ_0 is the extrapolation of the resistivity to 0 K, $\rho_T T$ is the roughly linear and independent contribution caused by temperature, $\rho_I C$ is the roughly linear contribution caused by solid solution impurities, and ρ_{other} represents the contributions from other scattering centers, such as dislocations and precipitates. Typical values of ρ_0 and α are listed in Table 3.1, along with the calculated resistivity at 100°C.

Experimentally determined resistivity versus temperature plots for a metal, an insulator, and a superconductor are shown in Figure 3.9. Superconductivity is the flow of electric current without resistance. It has been observed in certain metals, alloys, and ceramics at temperatures near absolute zero, and in some cases at temperatures hundreds of degrees above absolute zero. At low temperatures, all materials, other than superconductors, are insulators or metals. For pure metals, the resistivity increases rapidly with increasing temperature, whereas for insulators the resistivity decreases rapidly with increasing temperature [this was first observed by Michael Faraday (1791–1867) in 1833]. Semiconductors have resistivities intermediate between metals and insulators at room temperature. Instead of a square root dependence of temperature, the plot in Figure 3.9 reveals that, for a metal, there is proportionality with T at higher temperatures, and, at low temperatures, the resistivity is proportional to T^5! The latter is known as the Bloch-Gruneisen T^5 law.

From the above observations it is clear that some of the Drude assumptions are very wrong. Moreover, Drude's model cannot explain why one material acts as an insulator and another as a metal. This was all very confusing around Drude's time. Because solids contain a number of atoms and electrons with a similar density, why the large conductivity

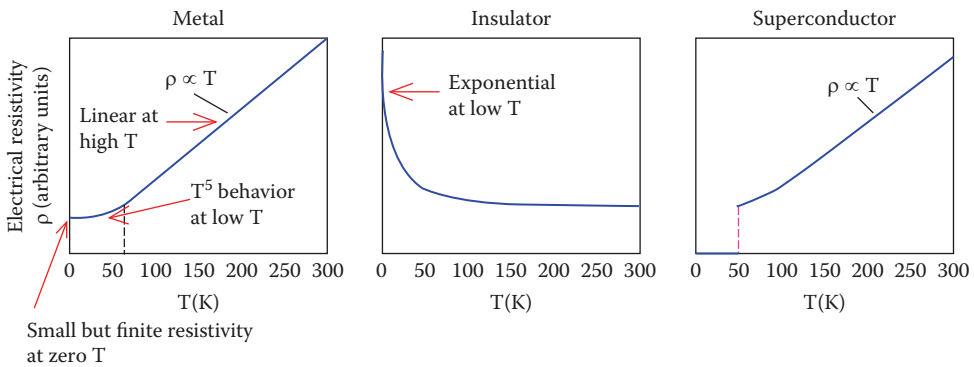


FIGURE 3.9 The three states of solid-state matter as defined by their electrical resistivity in the low temperature limit. The resistivity at low temperatures is finite for a metal, very large for an insulator, and zero for a superconductor.

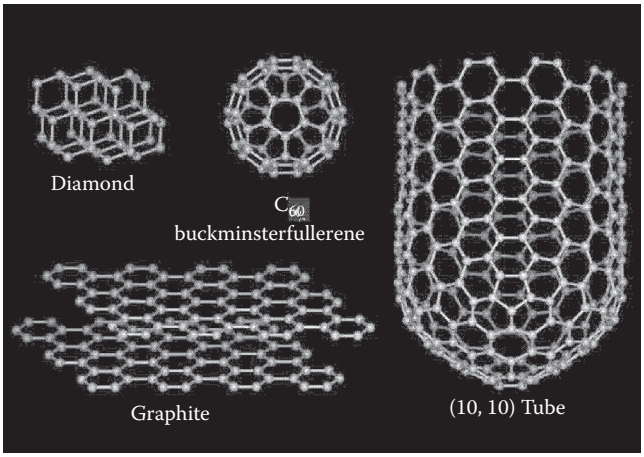


FIGURE 3.10 Carbon may act as an insulator (diamond), as a semimetal (graphite), and as a superconductor (buckminsterfullerene and carbon nanotube).

differences shown in Figure 3.2? More intriguing yet, in the case of carbon, the same material may act as a good insulator (diamond), as a semimetal (graphite), and even as a superconductor (buckminsterfullerenes and carbon nanotubes)* (see Figure 3.10). The answer, we will learn, is that electrons are *fermions*, and only electrons with energy on the order of a few k_BT contribute to the conduction process, at room temperature. A *fermion* is a particle, such as an electron, proton, or neutron, having half-integral spin and obeying statistical rules requiring that not more than one in a set of identical particles may occupy a particular quantum state. For fermions, the Boltzmann distribution must be replaced by a Fermi-Dirac distribution. Because of the wave nature of electrons and the exclusion principle (to be discussed below), the energy distribution of electrons in a metal does not even resemble a Maxwell-Boltzmann distribution. Moreover, the collision between ions and electrons cannot be pictured as that of two hard objects. Instead it involves the scattering of electron waves by lattice ions.

Drude AC Electrical Conductivity and Dielectric Functions

Drude AC Electrical Conductivity In the previous sections we considered the dc conductivity of metals; we now consider their ac electrical conductivity. We assume that the wavelength of the electromagnetic (EM) field is large compared with the electronic mean free path λ, so that electrons "see" a homogeneous field when moving between collisions. In Equation 3.4, we used Newton's second law to launch the change of momentum of a free electron in the presence of an applied dc electrical field. Here we use the same expression to derive the change of momentum of electrons in a time-harmonic electric field, given by:

$$\mathbf{E}(\omega,t) = \mathbf{E}(\omega)e^{-iwt} \tag{3.24}$$

The equation for the momentum per electron is:

$$\frac{d\mathbf{p}(\omega,t)}{dt} = -e\mathbf{E}(\omega,t) - \frac{\mathbf{p}(\omega,t)}{\tau} \tag{3.25}$$

where τ is the relaxation time or the mean free time a randomly picked electron travels before the next collision in a metal. Because $\frac{d\mathbf{p}(\omega,t)}{dt} = -i\omega\mathbf{p}(\omega,t)$, this expression may be rewritten as:

$$\begin{aligned}
\mathbf{p}(\omega,t) &= -\left[\frac{e\mathbf{E}(\omega,t)}{\frac{1}{\tau} - i\omega}\right] \text{ or also} \\
&= -\left[\frac{e\mathbf{E}(\omega,t)}{\frac{1}{\tau^2} + \omega^2}\right]\left(\frac{1}{\tau} + i\omega\right) \\
&= -\left[\frac{e\mathbf{E}(\omega)}{\frac{1}{\tau^2} + \omega^2}\right]\left(\frac{1}{\tau} + i\omega\right)\left(\cos\omega t - i\sin\omega t\right) \\
&= -\left[\frac{e\mathbf{E}(\omega)}{\frac{1}{\tau^2} - \omega^2}\right]\left(\frac{1}{\tau}\cos\omega t + \omega\sin\omega t\right) \\
&\quad + i\left(\omega\cos\omega t - \frac{1}{\tau}\sin\omega t\right)
\end{aligned} \tag{3.26}$$

We analyze this expression now for two different and important situations: 1) with $\omega \ll 1/\tau$, where the electrons can follow the changing electrical field, and 2) with $\omega \gg 1/\tau$, i.e., at very high frequencies

* The latter were not yet discovered at the time; if they had been, this would have led to even more consternation. How can the same material have all these different resistivities?

(optical frequencies we will see), where the electrons cannot follow the fast changing electrical field anymore.

1. With $\omega \ll 1/\tau$:

$$\mathbf{p}(\omega,t) = -\left[\frac{e\mathbf{E}(\omega)}{\frac{1}{\tau^2}}\right]\left(\frac{1}{\tau}\cos\omega t - i\frac{1}{\tau}\sin\omega t\right)$$

$$= -\left[\frac{e\mathbf{E}(\omega)}{\frac{1}{\tau}}\right]e^{-iwt} = -e\tau\mathbf{E}(\omega,t) \qquad (3.27)$$

   At these frequencies, $\mathbf{p}(\omega,t)$ is in phase with $\mathbf{E}(\omega,t)$.

2. With $\omega \gg 1/\tau$:

$$\mathbf{p}(\omega,t) = -\left[\frac{e\mathbf{E}(\omega)}{\omega^2}\right](\omega\sin\omega t + i\omega\cos\omega t)$$

$$= -\left[\frac{e\mathbf{E}(\omega)i\omega}{\omega^2}\right](\cos\omega t - i\sin\omega t)$$

$$= -\left[\frac{e\mathbf{E}(\omega)i\omega}{\omega^2}\right]e^{-iwt} = -\left[\frac{e\mathbf{E}(\omega,t)i}{\omega}\right]$$

$$= \frac{e\mathbf{E}(\omega,t)}{i\omega} \qquad (3.28)$$

At very high frequencies, where the electrons are too slow to follow the very fast changing electrical field vector, $\mathbf{p}(\omega,t)$ is out of phase with $\mathbf{E}(\omega,t)$; moreover, $\mathbf{p}(\omega,t)$ tends to zero. The transition between these two behaviors occurs when the ac frequency exceeds the collision frequency τ. With a typical value for τ of metal of about 10^{-14} s, that transition frequency is at roughly 10^{14} Hz ($\tau \sim 10^{-14}$ s, $v\lambda = c$ or $v = \frac{3.10^{10}\text{cm/s}}{5000\cdot10^{-8}\text{cm}} \approx 10^{14}$ Hz), corresponding to optical frequencies (light waves). The equations derived here apply for the ac behavior of metals as well their interaction with light (metal optics).

Relying on Equation 3.3, rendered in vector format as $\mathbf{J}(\omega) = -env_{dx}$, and given that $\mathbf{p}(t) = m_e\mathbf{v}(t)$, we can also write:

$$\mathbf{J}(\omega) = -\frac{en\mathbf{p}(\omega)}{m_e} = \frac{\left[\frac{e^2n}{m_e}\right]\mathbf{E}(\omega)}{\frac{1}{\tau} - i\omega} = \sigma(\omega)\mathbf{E}(\omega) \qquad (3.29)$$

so that the Drude complex AC conductivity is calculated as:

$$\sigma(\omega) = \frac{\sigma_0}{1 - i\omega\tau} \qquad (3.30)$$

which simplifies to the DC conductivity, i.e., σ_0 ($= \frac{ne^2\tau}{m_e}$, Equation 3.7), in the case of very low ac frequencies ($\omega \ll 1/\tau$).

The real and imaginary parts of the complex conductivity are:

$$\sigma' = \frac{\sigma_0}{1 + \omega^2\tau^2}(\text{Re}) \qquad (3.31)$$

and

$$\sigma'' = \frac{\sigma_0\omega\tau}{1 + \omega^2\tau^2}(\text{Im}) \qquad (3.32)$$

The real (σ') and imaginary (σ'') parts of the complex conductivity $\sigma(\omega)$ are plotted versus $\omega\tau$ in Figure 3.11. The maximum in σ'' is called the Drude peak and is characteristic for each metal.

We consider now what happens with the current and the conductivity as a function of frequency. At very low ac frequencies ($\omega \ll 1/\tau$), electrons have many collisions before the direction of the wave changes; this situation corresponds to the Ohm's law

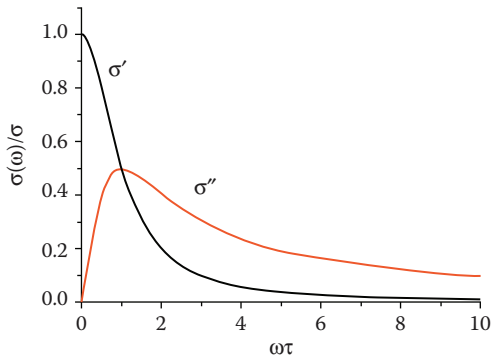


FIGURE 3.11 Frequency dependency of the real (σ') and imaginary (σ'') in parts of the conductivity. The maximum in σ'' at $\sigma''/\sigma_0 = 0.5$ is the Drude peak at $\omega\tau = 1$ and is a characteristic of a metal. The DC conductivity σ_0 is reached at $\omega\tau = 0$ ($\sigma''/\sigma_0 = 1$).

regime—**J** follows **E**—and σ is real. At very high frequencies ($\omega \gg 1/\tau$), electrons might have only one collision or less when the direction of the ac field is changed. In this case **J** is imaginary and out of phase with **E**, and σ is also imaginary. As we noted above, the transition from one regime to the other occurs at optical frequencies, and qualitatively we can say that with $\omega\tau \ll 1$, electrons are in phase and reirradiate, i.e., they are reflected and the metal appears shiny. With $\omega\tau \gg 1$, electrons are out of phase as they are too slow, there is less interaction, and we have transmission. More specifically, free electrons do not influence E-fields in metals with frequencies greater than that of visible light, or the electron gas is transparent in the UV range! No energy is absorbed from the field in this range, and no joule heating occurs.

Next, we will investigate how the permittivity or dielectric constant of a medium changes as a function of frequency of the applied electrical field (again from ac fields to electromagnetic radiation).

Dielectric Functions and an Introduction to Metal Optics The permittivity or dielectric constant of a medium describes how an electric field both affects and is affected by that medium and can be looked at as the quality of a material that allows it to store electrical charge. The dielectric constant ε or permittivity (*As/Vm*) is a materials-dependent "constant," and its frequency dependence defines the so-called dielectric function $\varepsilon(\omega)$. In the most general terms, dielectric functions are dependent on both frequency (ω) and wave-vector **k**: $\varepsilon(\omega,\mathbf{k})$. To understand the dielectric behavior and optical properties of metals better, we need to explain how plasma oscillations come about and analyze how the dielectric "constant" ε of a metal changes with frequency. A plasma, in general, is a medium with equal amounts of positive and negative charges, of which at least one charge type is mobile. In case of a plasma in a metal, the mobile charges are the free electrons, and these charges are balanced by the positive, immobile, metal ion cores. Drude's free electron gas (FEG), with an ac field applied, can exhibit collective longitudinal oscillations of the plasma because the displacement of all the electrons against the ion bodies of the material induces a dipole moment

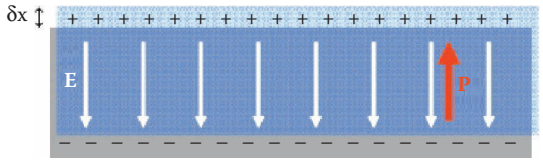


FIGURE 3.12 Drude bulk plasma oscillation in a metal film. **E** is the electrical field, and **P** is the polarization. The entire free electron gas is displaced over a small distance, δ*x*.

and an electric field opposing that displacement, as shown in Figure 3.12. In longitudinal oscillations, the displacement is in the same direction as the wave motion, as in sound waves. To appreciate better how oscillations of free electrons are induced by a time harmonic electrical field $\mathbf{E}(\omega,t) = \mathbf{E}(\omega)e^{-i\omega t}$, we rewrite Equation 3.25 for one dimension with $\mathbf{p}_x(\omega,t) = m_e\mathbf{v}_x(\omega,t) = m_e d_x/dt$ as:

$$m_e \frac{d^2x}{dt^2} + m_e\gamma\frac{dx}{dt} = -e\mathbf{E}(t) = -e\mathbf{E}(\omega)e^{-i\omega t} \qquad (3.33)$$

In the Drude model, electrons are free, and a damping factor γ comes about only because of electron scattering. The damping factor is related to the scattering constant as $1/\gamma = \tau$ (as we saw, typically ~10^{-14} s in a metal). The solution of Equation 3.33—the displacement *x* of the entire free electron gas—is then given as:

$$x(t) = \frac{e\mathbf{E}(\omega)}{m_e(\omega^2 + i\gamma\omega)} \qquad (3.34)$$

Polarization comes about as a result of small displacement of charges in an electrical field. The macroscopic polarization density **P** (C/m^2), illustrated in Figure 3.12, is a vector field that represents the density of permanent or induced electric dipole moments and is given by the product of the displacement $x(t)$ and the electron density ne(**P** = – nex) or:

$$\mathbf{P} = -\frac{ne^2\mathbf{E}(t)}{m_e(\omega^2 + i\gamma\omega)} \qquad (3.35)$$

The polarization **P** is also related to the electrical field as:

$$\mathbf{P}(\omega,t) = \chi_e\varepsilon_0\mathbf{E}(\omega,t) \qquad (3.36)$$

where χ_e is the electric susceptibility tensor of the material, a proportionality constant relating the electrical field **E** to the induced dielectric polarization

density **P**, and ε_0 is the permittivity of free space with a value of 8.854 pF/m. In the case of a linear homogeneous and isotropic material, χ_e becomes a scalar constant. The polarization **P** is further linked to the electric displacement field **D**, expressed in coulombs per square meter (C/m^2), and to the dielectric constant $\varepsilon(\omega,\mathbf{k})$ or permittivity (As/Vm), a materials-dependent constant, as:

$$\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} = (1 + \chi_e)\varepsilon_0\mathbf{E} = \varepsilon(\omega,\mathbf{k})\mathbf{E} \qquad (3.37)$$

In the expression $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$ we have separated the electric displacement **D** in its materials (**P**) and vacuum parts ($\varepsilon_0\mathbf{E}$). The permittivity of a material is usually given as a relative permittivity $\varepsilon_r(\omega)$ (also called the dielectric constant). The permittivity $\varepsilon(\omega,\mathbf{k})$ of a medium is an intensive parameter and is calculated by multiplying the relative permittivity by ε_0, or $\varepsilon(\omega) = \varepsilon_0\varepsilon_r(\omega)$. The permittivity of air is $\varepsilon_{air} = 8.876$ pF/m, so the relative permittivity of air is $\varepsilon_{r,air} = 1.0005$, and for vacuum it is 1 by definition. From Equation 3.37, it then also follows that $\varepsilon_r(\omega) = 1 + \chi_e$. The electric displacement field, **D**, is in turn related to the electric field **E** (in units of V/m) by the constitutive relation:

$$\mathbf{D} = \varepsilon(\omega,\mathbf{k})\,\mathbf{E} \qquad (3.38)$$

Using Equation 3.36 for the polarization **P**, we can rewrite Equation 3.37 as:

$$\mathbf{D} = \varepsilon(\omega,\mathbf{k})\mathbf{E} = \varepsilon_0\mathbf{E} + \mathbf{P} = \varepsilon_0\mathbf{E} - \frac{ne^2\mathbf{E}}{m_e(\omega^2 + i\gamma\omega)} \qquad (3.39)$$

From Equation 3.39 we arrive at the following expression for the complex dielectric function:

$$\varepsilon(\omega,\mathbf{k}) = \varepsilon_0\left[1 - \frac{ne^2}{m_e\varepsilon_0(\omega^2 + i\gamma\omega)}\right] = \varepsilon_0\left[1 - \frac{\omega_p^2}{(\omega^2 + i\gamma\omega)}\right] \qquad (3.40)$$

The plasma frequency ω_p in Equation 3.40 is defined as:

$$\omega_p = \sqrt{\frac{ne^2}{m_e\varepsilon_0}} \qquad (3.41)$$

where n is the density of electrons in the metal. To obtain $\varepsilon_r(\omega)$, the relative dielectric response of a material to electromagnetic waves at various frequencies, one divides by ε_0 because $\varepsilon(\omega,\mathbf{k})/\varepsilon_0 = \varepsilon_r(\omega,\mathbf{k})$. The corresponding real and imaginary components of the complex dielectric function are:

$$\varepsilon_r'(\omega,\mathbf{k}) = 1 - \frac{\omega_p^2\tau^2}{1 + \omega^2\tau^2} \qquad \text{Re} \qquad (3.42)$$

and

$$\varepsilon_r''(\omega,\mathbf{k}) = \frac{\omega_p^2\tau}{\omega(1 + \omega^2\tau^2)} \qquad \text{Im} \qquad (3.43)$$

In a metal, the damping term γ is just the electron collision rate, which is the inverse of the mean electron collision time, τ, i.e., $\gamma = \tau^{-1}$. As stated before, the collision rate can be quite rapid—tens of femtoseconds. But for optical frequencies (e.g., for $\lambda = 500$ nm, $\omega = 2\pi c/\lambda = 3.8 \times 10^{15}$ rad/s) $(\omega\tau)^2 >> 1$. Under this approximation, we find:

$$\varepsilon_r'(\omega,\mathbf{k}) \approx 1 - \frac{\omega_p^2}{\omega^2} \qquad (3.44)$$

and

$$\varepsilon_r''(\omega,\mathbf{k}) \approx \frac{\omega_p^2}{\omega^3\tau} = \frac{\omega_p^2\gamma}{\omega^3} \qquad (3.45)$$

This approximation may break down in the far-infrared spectral region, where damping can be significant. Note that damping (γ) is absolutely necessary to have an imaginary part of $\varepsilon_r(\omega)$. The dispersion of the real part of the dielectric function $\varepsilon_r'(\omega,\mathbf{k})$ (Equation 3.44) is plotted in Figure 3.13. The Drude model predicts a monotonous decrease of $\varepsilon_r'(\omega,\mathbf{k})$ for decreasing frequency, and experiments confirm

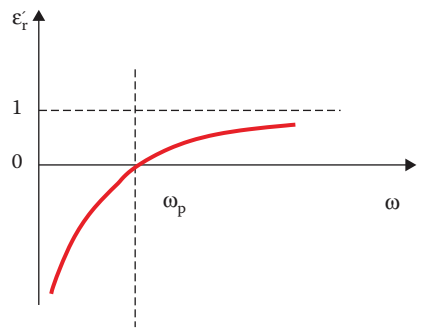


FIGURE 3.13 The dielectric function: ε_r' as a function of ω. The dielectric constant ε_r' becomes zero when $\omega = \omega_p$, and this supports free longitudinal collective modes for which all electrons oscillate in phase.

that $\varepsilon_r'(\omega,\mathbf{k})$ becomes zero when $\omega = \omega_p$. When $\varepsilon_r'(\omega)$ becomes zero at $\omega = \omega_p$, the material can support free longitudinal collective modes for which all electrons oscillate in phase. Such a collective charge oscillation is called a *plasmon*. Note that in the limit of γ going to 0, Equation 3.40 leads to the same result as shown in Equation 3.44 (no imaginary component to the dielectric constant at $\gamma = 0$). We come back to the field of plasmonics in Chapter 5 on photonics, and at that point we will not only analyze plots of $\varepsilon_r(\omega)$ for metals but also of $n(\omega)$ (refractive index), $\alpha(\omega)$ (absorption), and $R(\omega)$ (reflectance) for all types of materials.

With reference to Figure 3.12, we can derive the plasma frequency ω_p in yet another way. Let us look at the case where $\varepsilon_r'(\omega) = 0$ at $\omega = \omega_p$; because $\varepsilon_r'(\omega) = 1 + \chi_e'$, this means that at this frequency, $\chi_e' \approx -1$, and with the polarization, $\mathbf{P}(\omega,t) = \chi_e\varepsilon_0\mathbf{E}(\omega,t)$ **P** (Equation 3.36), this results in $\mathbf{P}(\omega,t) \approx -\varepsilon_0\mathbf{E}(\omega,t)$. The displacement δx of the electron gas of density n creates an electric field $\mathbf{E} = -\mathbf{P}/\varepsilon_0 = ne\delta x/\varepsilon_0$; this opposing field acts as the restoring force ($F_r = -eE$) on the electron gas. The equation of motion resulting from this restoring force F_r in such a simple electron gas oscillator is given as:

$$F_r = -eE = -\frac{ne^2\delta x}{\varepsilon_0} = m_e\frac{\partial^2(\delta x)}{\partial^2 t^2} \qquad (3.46)$$

This results in oscillations at the plasma frequency of $\omega_p = \sqrt{\dfrac{ne^2}{m_e\varepsilon_0}}$ (Equation 3.41).

The plasma frequency ω_p comes with a free space wavelength of:

$$\lambda_p = \frac{2\pi c}{\omega_p} \qquad (3.47)$$

In Table 3.2 we compare the plasma frequency and the free space wavelength for a generic metal and a generic semiconductor. Using Equation 3.41 we can now calculate ω_p for a variety of materials. For silver, for example, with a σ of 6.2×10^7 Ω-m, one obtains $\omega_p = 9.65 \times 10^{14}$ Hz (= 311 nm or 4 eV). The lower the resistivity (higher n), the higher the plasma frequency. Consequently, plasma frequencies ω_p are in the optical range (UV) for metals and in the THz to the infrared region for semiconductors and insulators. Metals reflect light in the visible region and are transparent at very high frequencies (UV and x-rays). With a free space wavelength less than λ_p, a wave propagates; when it is longer, the wave is reflected ($\omega > \omega_p$, or $\lambda < \lambda_p$).

Table 3.2 Plasma Frequency (ω_p) and Free Space Frequency (λ_p) for a Generic Semimetal and Semiconductor

Property	Metal	Semiconductor
Electron density: n, cm^{-3}	10^{22}	10^{18}
Plasma frequency: ω_p, Hz	5.7×10^{15}	5.7×10^{13}
Space frequency: λ_p, cm	3.3×10^{-5}	3.3×10^{-3}
Spectral range	UV	Infrared region

In a metal where both bound electrons [$\varepsilon_B(\omega)$] and conduction electrons [$\varepsilon_r(\omega)$] contribute an effective dielectric constant, $\varepsilon_{Eff}(\omega)$ must be defined as:

$$\varepsilon_{Eff}(\omega) = \varepsilon_B(\omega) + \varepsilon_r(\omega) \qquad (3.48)$$

where $\varepsilon_r(\omega)$ is the complex dielectric constant at $\omega\tau >> 1$, given by (see Equations 3.44 and 3.45):

$$\varepsilon_r(\omega) = \varepsilon_r'(\omega) + i\varepsilon_r''(\omega) = 1 - \frac{\omega_p^2}{\omega^2} + i\frac{\omega_p^2}{\omega^3\tau} \qquad (3.49)$$

Plasma oscillations can be excited in dielectric films as well. A dielectric material is a nonconducting substance whose bound charges are polarized under the influence of an externally applied electric field. In this case, although qualitatively the same effect as in a metal, the oscillations of the bound electrons are with respect to the immobile ions instead of with respect to the films boundary, and a restoring force related to the strength of the bond of the electrons to those ions must be introduced. In Chapter 5 on photonics, after introducing the Maxwell equations, the Drude plasmon model will be upgraded with the more complete Drude-Lorentz model that includes such a restoring force term. The restoring force pulls charges back in their equilibrium position. The Drude-Lorentz model without the restoring force reduces back to the Drude model, in which conduction electrons are not bound to atoms. There we will also calculate the effective dielectric constant $\varepsilon_{Eff}(\omega)$ (with bound and free electron contributions) of Equation 3.48.

Specific Heat Capacity of Metal and Insulators

Different materials require different amounts of heat to increase their temperature. Heat capacity per unit mass of a substance is known as specific heat, and as we will see, it is a measure of the number of degrees of freedom of the system. When using classical models, as in the case of electrical conductivity, things went wrong for calculations of the specific heat capacity of metals and insulators.

Based on the Drude assumption that electrons in a metal behave as a monatomic gas of N classical particles, they should be able to take up translational kinetic energy when the metal is heated. According to the equipartition principle of energy in a gas with N particles, the electron internal energy U, at temperature T, expressed per mole is:

$$\frac{U}{n} = \frac{3}{2}\frac{N}{n}k_B T = \frac{3}{2}N_A kT = \frac{3}{2}RT \qquad (3.50)$$

with n the number of moles in the system, N_A Avogadro's number, R the ideal gas constant, and k_B Boltzmann's constant. Remember also from thermodynamics that the first derivatives of the fundamental thermodynamic equations, in U (internal energy) or S (entropy), correspond to the intensive parameters T, P, and μ (chemical potential), whereas second derivatives correspond to important materials properties such as the molar heat capacity, or:

$$C_v = T\left(\frac{\partial s}{\partial T}\right)_v = \frac{T}{n}\left(\frac{\partial S}{\partial T}\right)_v = \frac{1}{n}\left(\frac{\partial Q}{\partial T}\right)_v \qquad (3.51)$$

This equation basically tells us that the molar heat capacity at constant volume is the quasistatic heat flux per mole required to produce a unit increase in the temperature for a system maintained at constant volume. Materials with high specific heat capacity (C_v) require more energy to reach a given temperature. In the case of a metal, one expects contributions to the heat capacity from both the lattice and the free electrons. The electronic contribution to the molar specific heat capacity at constant volume, C_v, from Equation 3.50, is:

$$C_{v,el} = \frac{\delta}{\delta T}\left(\frac{U}{n}\right)_v = \frac{3}{2}R \qquad (3.52)$$

or about 12.5 J/(mol·k).

Atoms in a lattice usually have no translational energy, and as temperature increases only vibrational energy increases. In the case of a monovalent metal, the lattice specific heat contribution, $C_{v,lat}$, equals $3R$. This follows from the fact that a classical atom oscillator has 3 degrees of freedom in vibration, or U = 3NkT. The vibration of a classical 1D harmonic oscillator is U = kT, with $\frac{1}{2}kT$ for kinetic and $\frac{1}{2}kT$ for potential energy (Figure 3.14). From the latter we expect $C_{v,lat}$ to be constant at $3R$ or 25 J/(mol·k), the so-called Dulong-Petit law. Experimentally, C_v, the sum of vibrational and electronic contributions ($C_v = C_{v,el} + C_{v,lat}$), is proportional to T^3 at low T and approaches a constant $3R$ at high T (see Figure 3.15). Diamond reaches the Dulong-Petit value 25 J/(mol·k) only at high temperature, whereas lead reaches the Dulong-Petit value at relatively low temperature.

From the above, at high temperature, electrons are expected to contribute to the lattice heat capacity for a total $C_v = C_{v,el} + C_{v,lat}$ or $\frac{9}{2}R$, which is obviously a different result from Dulong-Petit's $3R$! The total C_v for metals is found to be only slightly higher than that for insulators, which only feature lattice contributions. The question is then: where is the electronic contribution? The absence of a measurable contribution by electrons to the C_v was historically

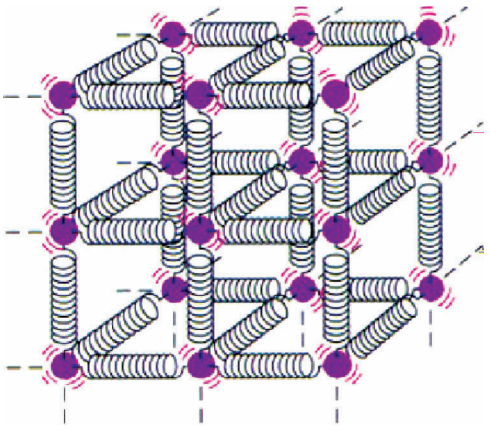


FIGURE 3.14 C_v: energy needed to raise T of 1 mol by 1 K at constant V, volume. The internal energy resides in vibrations of the solid constituents. The vibration of a classical 1D harmonic oscillator is U = kT. In a solid there are three perpendicular modes of vibration or three harmonic oscillators, and the total energy of a solid is thus U = 3NkT = 3RT.

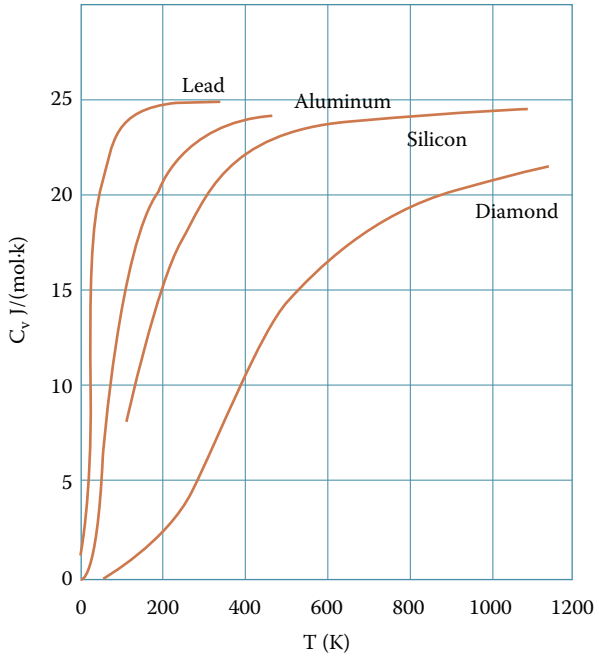


FIGURE 3.15 Dulong-Petit law with C_v = 3R or 25 J/(mol·k) at high temperatures.

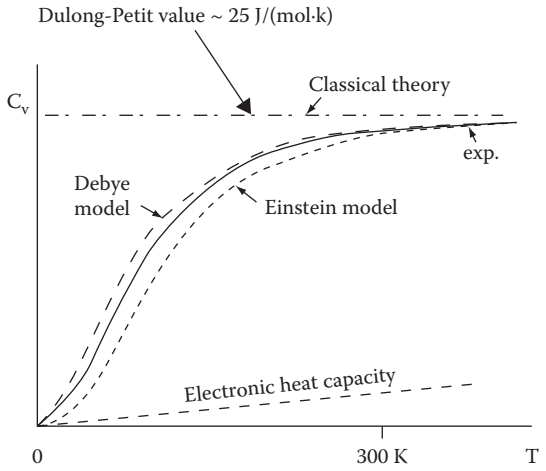


FIGURE 3.16 Schematic representation of the temperature dependence of the molar heat capacity, experimental and according to four models. For the Debye and Einstein models, see further below.

one of the major objections to the classical free electron model: because electrons are free to carry current, why would they not be free to absorb heat energy? Conduction electrons only contribute a small part of the heat capacity of metals, but, as we will see below, they are almost entirely responsible for the thermal conductivity.

In the above model, we used the principle of equipartition of the energy, which we borrowed from classical ideal gas theory. But, as Planck discovered in 1901, a vibrating system can only take up energy in quanta, the size of which is proportional to the vibration energy such that E = hν, where ν is frequency and *h* is the Planck constant. The chance that an atom vibrator can pick up an energy *h*ν is proportional to $e^{-h\nu/kT}$, and the average energy of a vibrating atom, at low temperature, may fall so low that the value of *kT* is now small compared with the size of the quantum *h*ν, and the probability factor $e^{-h\nu/kT}$ is dramatically decreased. Therefore, Dulong-Petit is incorrect at low temperatures as vibrations of successively lower frequencies fail to become excited. Atoms in a crystal do not obey Maxwellian statistics at low temperatures, and one needs to invoke the Einstein or Debye law instead. The reason that electrons do not contribute to the heat capacity at room temperature is, just like in the case of electrical conductivity, that they are fermions and cannot be treated as an ordinary Maxwell-Boltzmann gas. At room temperature, only the very few electrons in the so-called Maxwell-Boltzmann tail of the Fermi distribution can contribute to the heat capacity. Thus, at ordinary temperatures the electronic heat capacity is almost negligible, and it is the atomic vibration, i.e., the contribution of phonons, that dominates. In Figure 3.16 we show the experimental molar heat capacity as a function of temperature and the expectations for those values from classical theory and from the Einstein and Debye models, which are discussed at the end of this chapter. It is seen here that the electronic contribution $C_{v,el}$ varies linearly with temperature.

Thermal Conductivity

Heat conduction is the transfer of thermal energy from a hot body to a cold body. In a solid both electrons and phonons can move, and depending on the material involved, one or the other tends to dominate. We intuitively expect electrical and thermal conductivities somehow to be linked; from experience we know that, in general, good electrical conductors are also good thermal conductors. The connection between electrical and thermal conductivities for metals was first expressed in the Wiedemann-Franz law in 1853, suggesting that electrons carry thermal energy and electrical charge. Insulators are often transparent, and conducting

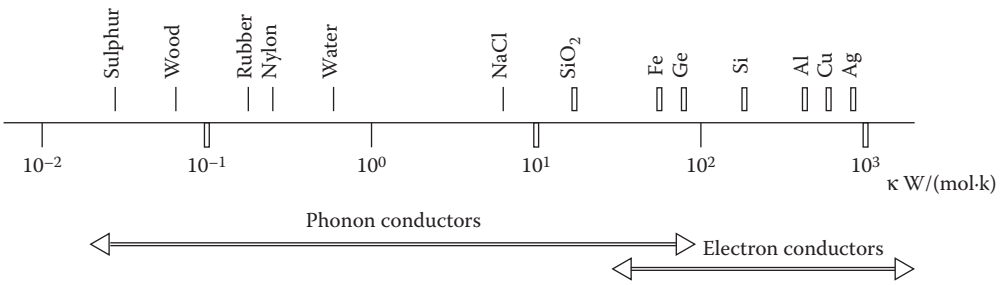


FIGURE 3.17 Phonon and electron conductors.

metals are reflecting and shiny when polished. Most metals are shiny because when light strikes a metal, the light is scattered by the moving electrons. For an exception, think about diamond, which is a transparent insulator but conducts heat better than aluminum or copper. For a glaring difference between heat and electrical conduction, consider that electrical conductivity of materials spans 25 orders of magnitude, whereas thermal conductivity only spans about four orders. In metals, heat is mostly transferred by electrons, but in electrical insulators, there are few free electrons, so the heat must be conducted in some other way, i.e., lattice vibrations or phonons. Thus, materials are divided into phonon conductors and electron conductors of heat as illustrated in Figure 3.17.

There is a major difference between heat conduction by electrons and by phonons; for phonons, the number changes with the temperature, but the energy is quantized, whereas for electrons, the number is fixed, but the energy varies. We will analyze now in more detail what that relation between electronic conductivity, σ, and thermal conductivity, κ is, and will discover that, when using classical models, the experimental results again cannot be properly explained.

Recall the electrical result $\mathbf{J} = \sigma\mathbf{E}$ (Equation 2.2) and also $\mathbf{J} = \sigma\mathbf{E} = \sigma\dfrac{dV}{dx}$. The thermal equivalent for this expression is Fourier's law for heat conduction, Q, which states that:

$$Q = \kappa A\frac{T_h - T_c}{L} = \kappa A\frac{dT}{dx} \tag{3.53}$$

with T_h the hot temperature and T_c the cold temperature, κ the thermal conductivity, L the thermal conduit path-length, and A its cross-sectional area. From the first law of thermodynamics (heat conservation), we know that the rate of heat conduction must equal the rate of change of energy storage:

$$\kappa\frac{\partial^2 T}{\partial x^2} = C_{v,el}\frac{\partial T}{\partial t} \tag{3.54}$$

The specific heat capacity, in general, is made up of a phonon and an electron term ($C_v = C_{v,el} + C_{v,lat}$). In most metals, the contribution of the electrons to heat conductivity greatly exceeds that of the phonons, and the phonon term can be neglected ($C_v = C_{v,el}$). Because of electrical neutrality, equal numbers of electrons move from hot to cold as the reverse, but their thermal energies are different.

Equations 3.53 and 3.54 only apply under the conditions that:

- t >> scattering mean free time of the energy carriers (τ)
- L >> scattering mean free path of the energy carriers (λ)

These conditions break down for applications involving thermal transport in small length/time scales, e.g., nanoelectronics, nanostructures, NEMS, ultrafast laser processing, etc. (see below and Volume III, Chapter 7 on scaling). For now, however, we are interested in the relation between electronic conductivity, σ, and thermal conductivity, κ, for somewhat larger systems.

To arrive at the expression for thermal conductivity, κ, we inspect Figure 3.18 and consider a metal bar [area A is a unit area (A = 1) here] with a temperature gradient dT/dx where we are interested in a small volume of material, with a length 2λ, with λ the mean free path between collisions. The idea is that an electron must have undergone a collision in this space and hence will have the energy/temperature of this location. To calculate the energy

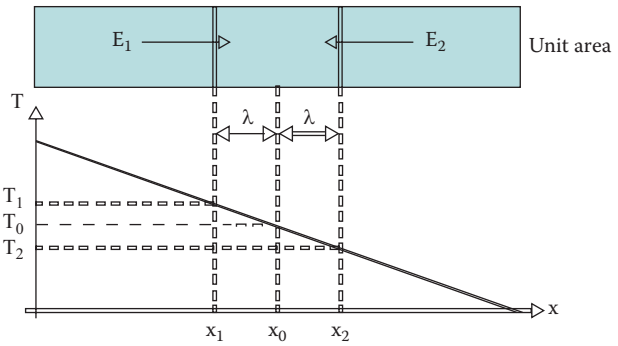


FIGURE 3.18 Fourier's Law for heat conduction: we are interested in a small volume of material, with a length 2λ, where λ is the mean free path between collisions with the lattice.

flowing per unit time per unit area from left to right (E_1), we multiply the number of electrons N crossing by the energy of one electron:

$$\text{energy} = \frac{1}{2}kT_1 = \frac{1}{2}k\left[T_0 + \lambda\left(-\frac{dT}{dx}\right)\right]$$

$$\text{number} = \frac{1}{6}nv_{dx} \qquad (3.55)$$

$$E_1 = \frac{nv_{dx}}{2}\frac{1}{2}k\left(T_0 - \lambda\frac{dT}{dx}\right)$$

with v_{dx} the drift velocity in the x-direction, and dT/dx a thermal gradient in the x-direction. We know that for charge neutrality the same number of electrons must flow in the opposite direction to maintain charge neutrality, but their energy per electron is lower.

Based on:

$$E_2 = \frac{nv_{dx}}{2}\frac{1}{2}k\left(T_0 + \lambda\frac{dT}{dx}\right) \qquad (3.56)$$

the thermal energy transferred per unit time per unit area is:

$$Q = E_1 - E_2 = -\frac{nv_{dx}}{2}k\lambda\frac{dT}{dx} \qquad (3.57)$$

Comparing this result with Equation 3.53 (with A = 1), we obtain:

$$\kappa = \frac{1}{2}nv_{dx}k\lambda \qquad (3.58)$$

This says that the thermal conductivity is larger if there are more electrons (n large), the electrons move faster (v_{dx} large), and they move more easily (large λ with fewer collisions). Because $N_Ak = R$ with R the ideal gas constant, $N/N_A = n$ (number of moles of electrons), and the electronic contribution to the molar specific heat capacity at constant volume, $C_{v,el}$, from Equation 3.52 is $3/2R$. The bulk heat conductivity of a solid per mole (n = 1) can then be rewritten as:

$$\kappa = \frac{1}{3}C_{v,el}v_{dx}\lambda \text{ or also } \frac{1}{3}C_{v,el}v_{dx}^2\tau \text{ sine } \lambda = v_{dx}\tau \quad (3.59)$$

As mentioned above, in most metals the conduction by electrons greatly exceeds that of the phonons; typically the phonon contribution at room temperature is only 1% of the electron contribution.

In 1853, long before Drude's time, Gustav Wiedemann and Rudolf Franz published a paper claiming that the ratio of thermal and electrical conductivities of all metals has almost the same value at a given temperature:

$$\frac{\kappa}{\sigma} = \frac{C_{v,el}mv_{dx}^2}{3ne^2} = \frac{3}{2}\left(\frac{k_B}{e}\right)^2 T \qquad (3.60)$$

calculated through Drude's application of classical gas law: $C_{v,el} = \frac{3}{2}nk_B$ and $\frac{1}{2}mv_{dx}^2 = \frac{3}{2}k_BT$. Ludwig Lorenz realized in 1872 that this ratio scaled linearly with temperature, and thus a Lorenz number, L, was defined as:

$$\frac{\kappa}{\sigma T} \equiv L \qquad (3.61)$$

which is very nearly constant for all metals (at room temperature and above). A typical value for L, say for Ag, is 2.31 10^{-8} WΩK^{-2} at 0°C and 2.37 10^{-8} WΩK^{-2} at 100°C (see Table 3.3).

Although Equation 3.61 is the correct relationship, we now know that the value for L calculated from it is wrong. However, Drude, using classical values for the electron velocity v_{dx} and heat capacity $C_{v,el}$, somehow got a number very close to the experimental value. But how lucky that Drude dude was: by a tremendous coincidence, the error in each term he made was about two orders of magnitude … in the opposite direction [the electronic $C_{v,el}$

TABLE 3.3 Some Typical Lorenz Numbers

Lorenz Number L in 10^{-8} WΩK^2		
Metal	273 K	373 K
Ag	2.31	2.37
Au	2.35	2.40
Cd	2.42	2.43
Cu	2.23	2.33
Ir	2.49	2.49
Mo	2.61	2.79
Pb	2.47	2.56
Pt	2.51	2.60
Sn	2.52	2.49
W	3.04	3.20
Zn	2.31	2.33

is 100 times smaller than the classical prediction, but $(v_{dx})^2$ is 100 times larger]. So the classical Drude model gives the prediction:

$$L_{Drude} = 1.12\,10^{-8}\,W\Omega K^{-2} \tag{3.62}$$

But in Drude's original paper, he inserted also an erroneous factor of two, as a result of a mistake in the calculation of the electrical conductivity. So he originally reported:

$$L = 2.24 \times 10^{-8}\,W\Omega K^{-2} \tag{3.63}$$

The correct value for the Lorenz number, L, derived from quantum mechanics is $\frac{\pi^2 k_B^2}{3e^2}$ or 2.45×10^{-8} WΩK^2 (see Equation 3.340 below). So although Drude's predicted electronic heat capacity was far too high (by a factor of 100!), his prediction of L made the free electron gas (FEG) model seem more impressive than it really was and led to a general acceptance of the model.

Hall Coefficients

Conductivity measurements do not yield information about the sign of charge carriers; for this we need the Hall effect. The Hall effect uses current and a magnetic field to determine mobility and the sign of charge carriers. The Hall effect, discovered in 1879 by American physics graduate student (!) Edwin Hall, is simple to understand.[1] With reference to Figure 3.19, consider a fairly strong magnetic field **B** (~2000 Gauss) applied perpendicular to a thin metal film (thickness is d and width is w) carrying a current I. The magnetic Lorentz force is

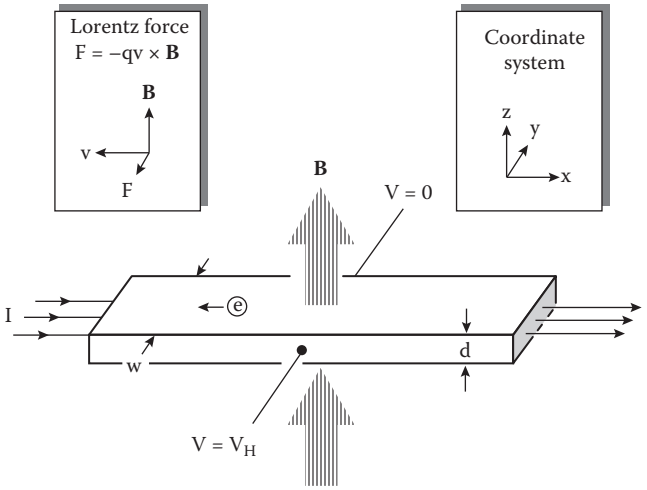


FIGURE 3.19 The Hall effect. Magnetic field **B** is applied perpendicular to a thin metal film sample carrying current I. V_H is the Hall voltage.

normal to both the direction of the electron motion and the magnetic field. The path of the charge carriers shifts as a result of this Lorentz* force, and, as a consequence, the Hall voltage, $V_H = E_y w$ (where E_y is the Hall field strength), builds up until it prevents any further transverse displacement of electrons. In other words, the Hall field increases until it is equal to and opposite of the Lorentz force. The orientation of the fields and sample is illustrated in Figure 3.19.

In mathematical terms, the above translates as follows. With both electric and magnetic fields present, a charge carrier with charge q experiences a Lorentz force in the lateral direction:

$$\mathbf{F}_L = q(\mathbf{v}_d \times \mathbf{B}_z) \tag{3.64}$$

We do use the drift velocity v_d of the carriers because the other velocities (and the forces caused by these components) cancel to zero on average. Note that instead of the usual term "electron," the term "charge carrier" is used here because in principle an electrical current could also be carried by charged particles other than electrons, e.g., positively charged ions or holes (missing electrons). The vector product in Equation 3.64 ensures that the Lorentz force is perpendicular to $\mathbf{v}_d$ and $\mathbf{B}_z$. For the geometry assumed in Figure 3.19, the Lorentz force $\mathbf{F}_L$ has only a component in the y-direction, and we can use a

* This is the Lorentz we encounter again in Chapter 5 (Figure 5.50), not the Lorenz from the Lorenz number L.

scalar equation for the Lorentz force: $F_L = -qv_dB_z$ or with Equation 3.11 (where we replaced μ_e by μ_q for generality):

$$\mathbf{F}_L = -q\mu_q E_x \mathbf{B}_z \tag{3.65}$$

As more and more carriers are deflected by the magnetic force, they accumulate on one side of the thin-film conductor, and this accumulation of charge carriers leads to the "Hall field" $\mathbf{E}_y$ that imparts a force opposite to the Lorentz force. The force from that electrical field $\mathbf{E}_y$ in the y-direction (which is of course $q \cdot \mathbf{E}_y$) must be equal to the Lorentz force with opposite signs. This way we obtain:

$$\begin{aligned} F_L &= -q\mu_q E_x B_z = qE_y \text{ or} \\ F_L &= -\mu_q E_x B_z = E_y \end{aligned} \tag{3.66}$$

The Hall voltage V_H now is simply the field in y-direction multiplied by the dimension w in the y-direction (=wE_y). It is clear then that the (easily measured) Hall voltage is a direct measure of the mobility μ of the carriers involved, and that its sign or polarity will change if the sign of the charges changes.

It is customary to define a Hall coefficient, R_H, for a given material as:

$$R_H \equiv \frac{E_y}{J_x B_z} \tag{3.67}$$

In other words, we expect that the Hall voltage, wE_y, will be proportional to the current density $\mathbf{J}_x$ and the magnetic field strength $\mathbf{B}$, which are, after all, the main experimental parameters (besides the trivial dimensions of the specimen):

$$\mathbf{E}_y = R_H \mathbf{B}_z J_x \tag{3.68}$$

Using Equations 3.10, 3.66, and 3.67, as well as $\mathbf{J}_x = \sigma \mathbf{E}_x$, we calculate R_H as:

$$R_H = -\frac{-\mu_q E_x B_z}{\sigma E_x B_z} = -\frac{\mu}{\sigma} = -\frac{\mu}{qn\mu} = -\frac{1}{qn} \tag{3.69}$$

When we calculate R_H from our measurements and assume $|q| = e$ (which Hall at the time did not know!), we can find n, the charge density. Also, the sign of V_H and thus of R_H tells us the sign of q! If R_H is negative, the predominant carriers are electrons; if positive, they are holes. If both types of carriers are present, the Hall field will change its polarity depending on the majority carrier (Figure 3.20). Also from Equation 3.69 one sees that the lower the carrier density, the higher R_H and thus the higher V_H; this is a key to some very sensitive magnetic field sensors.

For most metals the Hall constant is negative because electrons are majority carriers. But for Be and Zn, for example, there is a band overlap with dominant conduction by holes in the first band and fewer electrons in the second. As a result, R_H is positive for these metals. Of course energy bands were not heard of yet, and the results of a positive R_H were baffling at the time: how can we have $q > 0$ (even for metals!)? The Hall coefficient changes sign with the sign of the charge carrier and therefore provides an important method for investigating the electronic structure of the solid state. In particular, the positive Hall coefficients exhibited by metals such as magnesium and aluminum are a clear indication that a naive picture of a sea of conduction electrons is inappropriate because the majority carriers are clearly positively charged (and are, in fact, holes). The discrepancies between the FEG predictions and experiments nearly vanish when liquid metals are compared (see Table 3.4). This

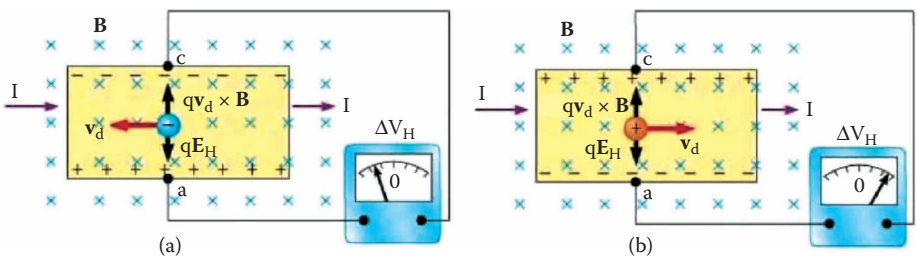


FIGURE 3.20 Electron and hole charge carriers in the Hall effect. When the charge carriers are negative, the upper edge of the conductor becomes negatively charged (a). When the charge carriers are positive, the upper edge becomes positively charged (b).

TABLE 3.4 Discrepancies between the FEG Predictions for R_H and Experiments Nearly Vanish When Liquid Metals Are Compared*

		R_H (10^{-11} m^3/As)		
Metal	n_0	Solid	Liquid	FEG Value
Na	1	–25	–25.5	–25.5
Cu	1	–5.5	–8.25	–8.25
Ag	1	–9.0	–12.0	–12.0
Au	1	–7.2	–11.8	–11.8
Be	2	+24.4	–2.6	–2.53
Zn	2	+3.3	–5	–5.1
Al	2	–3.5	–3.9	–3.9

*This reveals clearly that the source of these discrepancies lies in the electron-lattice interaction. Notice positive R_H for Be and Zn.

reveals clearly that the source of these discrepancies lies in the electron-lattice interaction of a solid.

The Hall effect "oddities" will be properly explained once we have introduced quantum physics. We will also see that under special conditions of extremely low temperature, high magnetic field, and two-dimensional electronic systems (2D electron gas, in which the electrons are confined to move in planes), the Hall resistance R_H is quantized and increases as a series of steps with increasing magnetic field. These discrete energy levels are called Landau levels.

Blackbody Radiation

Planck first discovered the discontinuous behavior that characterizes the atomic world in his analysis of the light spectra emitted from blackbodies in 1900. All substances, with thermal energy, radiate EM waves, and the radiation emanating from solids consists of a continuous spectrum of wavelengths. A blackbody (also cavity radiation) is a hypothetical object that is a perfect absorber or perfect emitter of radiation (Figure 3.21).

"Blackbody" is an unfortunate name as the ideal radiating/absorbing body does not have to be black; stars and planets, to a rough approximation, are blackbodies! When a blackbody is heated, it first feels warm and then glows red or white hot, depending on the temperature. A typical spectrum of the radiated light intensity, brightness, or emittance of a blackbody is shown in Figure 3.22. Intensity is a measure of how much energy is emitted from an object per unit surface area per unit time at a given wavelength

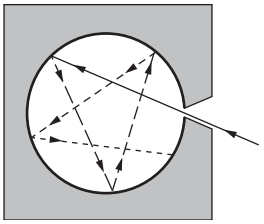


FIGURE 3.21 Ideal blackbody (also cavity radiation). "Blackbody radiation" or "cavity radiation" refers to an object or system that absorbs all radiation incident on it and reradiates energy that is characteristic of this radiating system only, not dependent on the type of radiation that is incident on it. The radiated energy can be considered to be produced by standing wave or resonant modes of the cavity, which is radiating.

and in a particular direction. A blackbody of temperature T emits a continuous spectrum peaking at λ_{max}. At very short and very long wavelengths there is little light intensity, with most energy radiated in some middle range frequencies. As the body gets hotter, the peak of the spectrum shifts toward shorter wavelengths (higher frequencies), but there is always a cutoff at very high frequencies.

It was experimentally observed that the brightness peak position shifted with temperature as:

$$T\lambda_{max} = \text{constant} = 2.898 \times 10^{-3}\ \text{m·K} \quad (3.70)$$

This is known as the Wien displacement law (see Figure 3.22). It was also known that the total energy, E, could be represented as the Stefan-Boltzmann law:

$$E = \sigma T^4 \quad (3.71)$$

where the constant $\sigma = 5.6704 \times 10^{-8}$ W/m²·K⁴. Classical interpretation predicted something

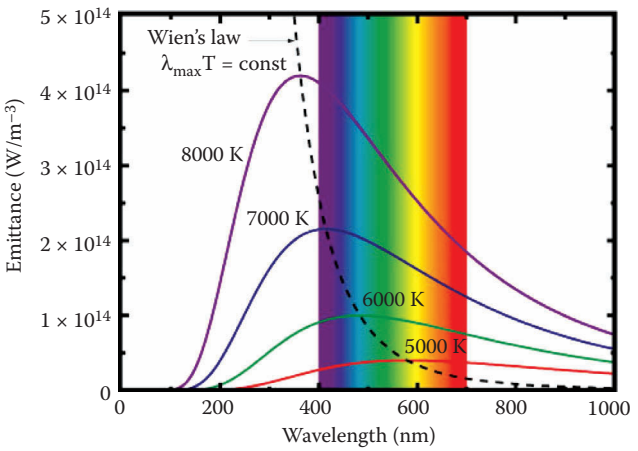


FIGURE 3.22 Blackbody radiation spectra at four different temperatures.

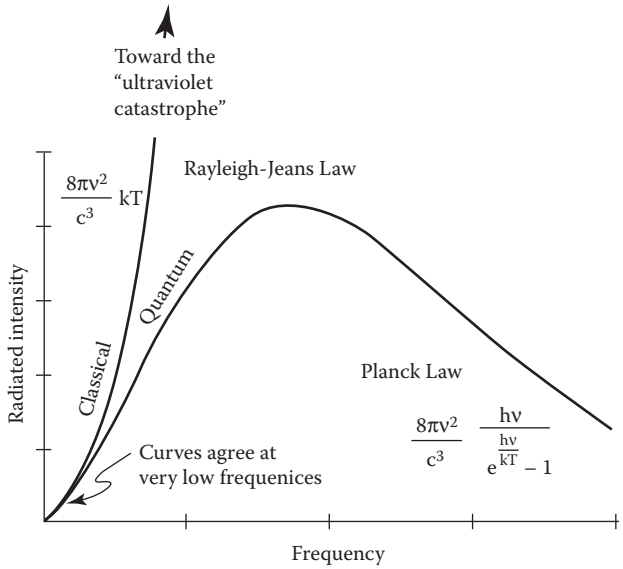


FIGURE 3.23 Planck and Rayleigh-Jeans models for blackbody radiation.

altogether different; in the classical Rayleigh-Jeans model, instead of a peak in the blackbody radiation and a falling away to zero at zero wavelength, the measurements were predicted to go off scale at the short wavelength end as shown in Figure 3.23.

Here is how the British physicists Lord Rayleigh and Jeans derived their model. They interpreted the blackbody radiation coming from a solid as electromagnetic radiation from oscillators that vibrate at every possible wavelength λ. In Figure 3.23 the radiated intensity, $E(\lambda)$, in $Js^{-1}\ m^{-3}$, is the energy distribution, i.e., the energy, E, at each λ. At equilibrium, the mean energy of all oscillators at temperature T is kT. The energy E_λ in a small interval $d\lambda$ is then given by:

$$E_\lambda\, d\lambda = kTdn \tag{3.72}$$

where dn is the fraction of oscillators at the "average energy" in the $d\lambda$ interval. From the Maxwell equations (see Chapter 5) one derives:

$$\frac{dn}{d\lambda} = \frac{8\pi}{\lambda^4} \tag{3.73}$$

Substituting this relation in Equation 3.72 leads to:

$$E_\lambda d\lambda = kT\left[\frac{8\pi}{\lambda^4}\right]d\lambda$$

or also since $\nu = c/\lambda$ (3.74)

$$E_\lambda d\lambda = kT\left[\frac{8\pi\nu^2}{c^3}\right]d\nu$$

The term in brackets $\left[\frac{8\pi\nu^2}{c^3}\right]$ is the number of modes per unit frequency per unit volume.

The amount of radiation emitted in a given frequency range should be proportional to the number of modes in that range. The best of classical physics suggested that all modes had an equal chance of being produced, and that the number of modes increased proportionally to the square of the frequency. Equation 3.74 works at long wavelengths (see Figure 3.23) but fails at short wavelengths. This failure at short wavelengths is called the ultraviolet catastrophe; as λ decreases, E_λ goes to $+\infty$. Even at very low T, the exponential curves in Figure 3.23 would have a huge value for visible light. Objects would be visible in the dark. The predicted continual increase in radiated energy with frequency (dubbed the "ultraviolet catastrophe") did not happen. Nature knew better.

The UV catastrophe attracted the attention of many physicists at the end of the nineteenth century, including Max Planck. Planck took the revolutionary step that led to quantum mechanics. He concluded that the available amount of energy could only be divided into a finite number of pieces among the atom oscillators in the cavity walls, and the energy of such a piece of radiation must be related to its frequency according to a new extremely important equation:

$$E = h\nu \tag{3.75}$$

where h was a new constant now called the Planck constant. The quantization implies that a photon of blue light of given frequency or wavelength will always have the same size quantum of energy. For example, a photon of blue light of wavelength 450 nm will always have 2.76 eV of energy. It occurs in quantized chunks of 2.76 eV, and you cannot have half a photon of blue light—it always occurs in precisely the same-sized energy chunks. Planck showed that the intensity I of radiation from a blackbody could be described by the function (now known as the Planck function):

$$I(\lambda,T) = \frac{2hc^2}{\lambda^5}\frac{1}{\left(e^{\frac{1}{\lambda kT}}-1\right)} \tag{3.76}$$

This Planck function was initially found empirically by trial and error but later derived by assuming the Planck constant and a Boltzmann distribution. Differentiating the Planck function with respect to wavelength one derives Wien's displacement law—the wavelength of the maximum:

$$\lambda_{max} = \frac{hc}{4.965kT} \qquad (3.77)$$

The constant hc/4.965kT agrees with the 2.898×10^{-3} m·K experimental value from Equation 3.70. Integration of the Planck function with respect to wavelength over all possible directions results in the total energy emitted per unit area per unit time from the surface of a blackbody (or absorbed per unit area per unit time by a body in a blackbody radiation field). This gives us back the Stefan-Boltzmann law (Equation 3.71):

$$E = \int_0^\infty E_\lambda(T) = \frac{2\pi^5 k^4}{15c^2h^3} T^4 = \sigma T^4 \qquad (3.78)$$

which also enables us to verify that the constant σ is indeed 5.6704×10^{-8} W/m²·K⁴.

Planck, interestingly, never appreciated how far removed from classical physics his work really was. He spent most of his life trying to reconcile the new ideas with classical thermodynamics. The Planck constant *h* was a bit like an "uninvited guest" at a dinner table; no one was comfortable with this new guest. But today we know that discontinuities in the nanoworld are meted out in units based on this Planck constant. This constant and its particular magnitude constitute one of the great mysteries in nature. It is the underlying reason for the perceived weirdness of the nanoworld, the existence of a "least thing that can happen" quantity—a quantum. The ubiquitous occurrence of discontinuities in the nanoworld constantly upsets our commonsense understanding of the apparent continuity of the macroscopic world.

Light as a Stream of Particles

Photoelectric Effect

The photoelectric effect, discovered by chance, by Heinrich Hertz and his student Hallwachs in 1888,

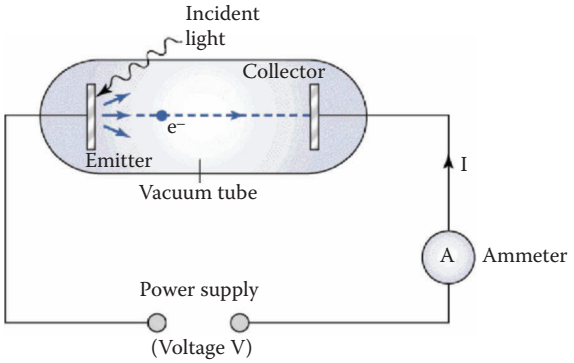


FIGURE 3.24 Experimental setup for studying the photoelectric effect.

is a process whereby light falling on negatively charged Zn, in an evacuated vessel, knocks electrons out of the surface as illustrated in Figure 3.24. The details of the photoelectric effect were in direct contradiction to the expectations of very well-developed classical physics of light waves, and the correct explanation by Einstein in 1905 marked one of the major steps toward quantum theory. Classical light wave theory predicts that the electrons in the metal will absorb radiation energy continuously. Once an electron has absorbed energy in excess of its binding energy, it will be ejected, and one would expect that a higher intensity would increase the chance that electrons are ejected.

As illustrated in Figure 3.24, to test this model we introduce a metal collector plate that collects electrons and a circuit that may be closed to measure the current *I*. In this setup, a retarding or stopping potential, V_0, may be applied to determine the kinetic energy of the electrons ($eV_0 = 1/2\ mv^2$). The classical wave model implies that the stopping voltage V_0 must be proportional to the intensity. Increasing frequency ν should not matter much, perhaps only causing a small decrease in current *I*, as a result of the rapid wave oscillations at high frequencies. With low intensity light, there should be a time delay to build up enough energy before current starts to flow.

The results were unexpected; no electrons were ejected, regardless of the intensity of the light, unless the frequency exceeded a certain threshold characteristic of the bombarded metal (red light did not cause the ejection of electrons, no matter what the intensity). The electrons were emitted

immediately—no time lag. Increasing the intensity of the light increased the number of photoelectrons but not their maximum kinetic energy (a weak violet light would eject only a few electrons, but their maximum kinetic energies were greater than those for intense light of longer wavelengths). The kinetic energy of the ejected electrons varied linearly with the frequency of the incident radiation but was independent of the intensity, or:

$$eV_0(\text{ = kinetic energy of the electrons}) = h\nu - \Phi \tag{3.79}$$

where Φ, the photoelectric work function, is the energy lost by a surface when an electron is freeing itself from its environment. This cannot be explained by the Maxwell equations.

If the charge of the electron is known, a plot of retarding or stopping voltage versus frequency of incident light, shown for three different metals in Figure 3.25, may yield a value for the Planck constant *h*. The electron charge was determined by Robert Millikan in 1909, and with that value and the slope of the lines in Figure 3.25, a value for *h* of 6.626×10^{-34} J·s was calculated, identical to the one derived from the hydrogen atom spectrum and blackbody radiation (see above). The intercept with the frequency axis (at kinetic energy zero) is the threshold frequency, ν_0, and the stop or retarding voltage axis intercept is the binding energy ($-\Phi$) of the electron.

The photoelectric phenomenon could not be understood without the concept of a light particle, i.e., a quantum amount of light energy for a particular frequency. Einstein's paper explaining the photoelectric effect was one of the earliest applications of quantum theory and a major step in its establishment. The remarkable fact that the ejection energy was independent of the total energy of illumination showed that the interaction must be like that of a particle that gave all of its energy to the electron! This fit in well with Planck's hypothesis that light in the blackbody radiation experiment could exist only in discrete bundles with energy. In quantum theory, the frequency, ν, of the light determines the energy, *E*, of the photons and $E = h\nu$, where *h* is Planck's constant ($h = 6.626069 \times 10^{-34}$ J·s) (Figure 3.26).

This assumption explains quantitatively all the observations associated with the photoelectric effects. A photon hits an electron and is absorbed. The energy of the emitted electron is given by the energy of the photon minus the energy needed to release the electron from the surface. This explains the observance of a threshold value below which no electrons are emitted. Thus, it depends on the frequency of light falling on the surface but not on its intensity. It also explains why there is no time lag; a photon hits an electron, is absorbed by the electron, and the electron leaves. Higher intensity light contains more photons, and so it will knock out more electrons. However, if the frequency of the light is such that a single photon is not energetic enough to release an electron from the surface, then none will be ejected no matter how intense the light. Gilbert N. Lewis in 1926 called Einstein's light particles photons. Just as the word *photon* highlights the particle aspect of an electron, the word *graviton* emphasizes

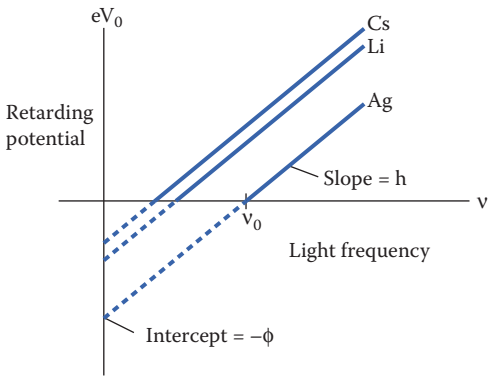


FIGURE 3.25 A plot of retarding or stopping voltage versus frequency of incident light. Slope is the Planck constant *h*. The intercept with the frequency axis (at kinetic energy zero) is the threshold frequency, ν_0, and the stop voltage axis intercept is the binding energy.

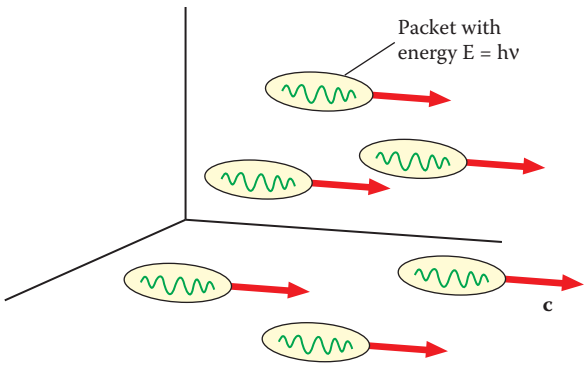


FIGURE 3.26 Light as energy packets with an energy $E = h\nu$.

the particle aspect of gravity, and *gluon* that of the strong nuclear force.

The young Einstein, in 1905, was the first scientist to interpret Planck's work as more than a mathematical trick and took the quantization of light ($E = h\nu$) for physical reality. He gave the uninvited dinner guest—the Planck constant h—a place at the quantum mechanics dinner table. What Einstein proposed here was much more audacious than the mathematical derivations by Planck to explain away the UV catastrophe. For a long time the science of optics had hesitated between Newton's corpuscular hypothesis and Huygens' wave theory. By the beginning of the nineteenth century the wave theory had become the dominant theory, largely because of the persistence of Augustin Fresnel (1788–1827), who described diffraction mathematically, and James Clerk Maxwell (1831–1879), who introduced the famous Maxwell equations in 1864 (Chapter 5). No wonder Einstein wrote at that time to one of his friends: "I have just published a paper about light, but I am sure nobody will understand it." Einstein reintroduced a modified form of the old corpuscular theory of light, which had been supported by Newton but which was long abandoned.

The particle nature of light was hard to swallow at the time, and it indeed still is. Remember the diffraction of x-rays on a crystal described in Chapter 2. Diffraction is something that happens with waves, not with particles. Einstein's light particles also negated the issue of "ether," a medium required for wave propagation as in the case of sound waves (sound waves cannot propagate in a vacuum); light is perfectly happy traveling in a vacuum.

How can we reconcile this duality of a photon as both wave- and particle-like in nature? We will have to learn to think of the wave as the probability of finding the particle. For example, if we know the momentum of the particle exactly, we cannot say where it is, only where it is likely to be [the Heisenberg uncertainty principle (HUP)].

Compton Scattering

The first strong support for the quantum nature of light came from monochromatic x-ray scattering on a graphite block. In 1922, Arthur Compton

FIGURE 3.27 Arthur Harry Compton (1892–1962).

(Figure 3.27), at Washington University in St. Louis, saw that the wavelength of x-rays increases on scatterings off graphite, depending on the angle (Figure 3.28). This effect cannot be explained using wave theory of x-rays.

As first explained by Compton in 1923, a photon can lose part of its energy and momentum on scatterings with electrons in the graphite, and the resulting energy loss (or change in wavelength, $\Delta\lambda$) can be calculated from the scattering angle θ. The result is that some of the scattered radiation has a smaller frequency (longer wavelength) than the incident

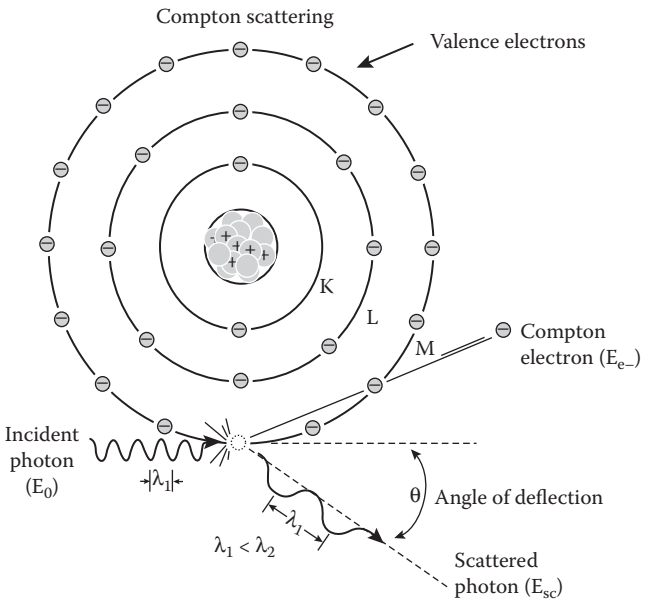


FIGURE 3.28 Compton effect. An incoming photon (E_0) can inelastically scatter from an electron and lose energy, resulting in an outgoing scattered photon (E_{sc}) with lower energy ($E_{sc} < E_0$).

radiation. This change in wavelength depends on the angle through which the radiation is scattered, and Compton concluded that an x-ray photon—or light in general—possesses momentum and thus behaves as a particle. Obviously, this was a big boost for the theory of light as composed of quanta. The three experiments that made the quantum revolution—blackbody radiation, the photoelectric effect, and the Compton effect—all indicate that light consists of particles.

Quantum Mechanics to the Rescue

Introduction

Invoking quantum mechanics we can solve the many problems with classical theories we exposed in the previous sections. Central to quantum mechanics is Schrödinger's equation, but before introducing Schrödinger's equation, we must put down some more foundations. We start with Kelvin's and Thomson's plum pudding atom model, then review Rutherford's and Bohr's improved orbital atom models, emphasize the importance of the Balmer's hydrogen emission lines in the discovery of the inner structure of an atom, and then we get baffled by de Broglie's matter waves and Heisenberg's uncertainty principle (HUP). There are four principal representations of quantum mechanics: Dirac's Hamiltonian and quantum algebra representation; the matrix representation of Born, Heisenberg, and Jordan; Schrödinger's wave equations; and Feynman's sum-over-histories approach; the latter constitutes a fundamentally new way of looking at quantum theory. With the Schrödinger formalism of quantum mechanics we tackle the band model and revisit electrical and thermal conductivity and heat capacity.

Bohr's and Rutherford's Atom

The Greek philosopher Leucippus of Miletus, who lived around 400 BC, first proposed atomic theory of matter. His disciple, Democritus of Abdera, concluded that infinite divisibility of a substance belongs only in the imaginary world of mathematics and further developed his mentor's atomic theory (Figure 3.29).

FIGURE 3.29 Democritus.

Democritus suggested that all things are "composed of minute, invisible indestructible particles of pure matter." According to the ancient Greeks, "atomos"* or atoms were all made of the same basic material, but atoms of different elements had different sizes and shapes. The sizes, shapes, and arrangements of a material's atoms determined the material's properties. For example, sour-tasting substances were believed to contain atoms with jagged edges, whereas round atoms made substances oily. It was further believed that there were four elements that all things were made of: earth, air, fire, and water. This basic theory remained unchanged until the nineteenth century, when it first became possible to test the theory with more sophisticated experiments. Lord Kelvin and J.J. Thomson developed a "raisin cake" model of the atom (also called the plum pudding model, 1897), which incorporated Thomson's negatively charged electrons as the raisins in a positively charged cake (Figure 3.30).

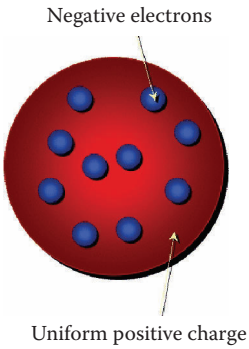


FIGURE 3.30 Kelvin's and Thomson's raisin cake.

* Atomos in Greek means "unbreakable."

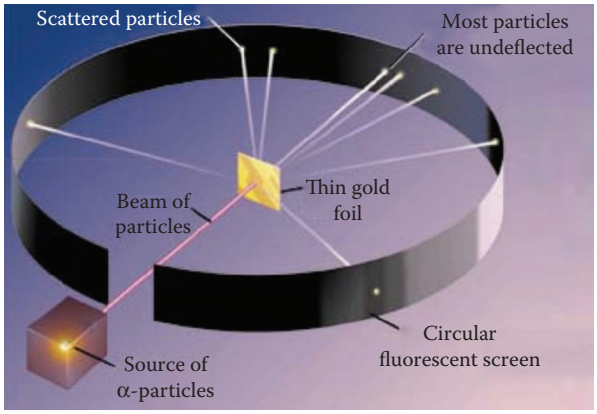


FIGURE 3.31 The Rutherford experiment. Structure of atom experiments.

To check the Kelvin and Thomson model, Geiger and Marsden, working in Rutherford's lab in 1911, directed a narrow beam of alpha particles, double-ionized helium atoms, of known energy onto a thin gold foil. A ZnS scintillation screen was used to record the striking alpha particles. Most particles went through the gold film undeflected, and some were deflected at small angles; however, they found that, once in a while, for about 1% of particles, the α-particles were scattered backward by the target (Figure 3.31).

For a moving alpha particle to be scattered through a large angle, a considerable repulsive force must be exerted. To explain the backscattering, Rutherford proposed that the positive charge in a gold atom must be concentrated in a small region. Rutherford proposed in this way the first realistic model of the atom: he concentrated 99.99% of the mass of the atom in the nucleus, which is only 10^{-15} m across (Figure 3.32). In other words, the atom (and therefore matter in general) is composed of 99.9999999999999% empty space. The proportion of nucleus to total atom size is obviously not drawn to scale in Figure 3.32.

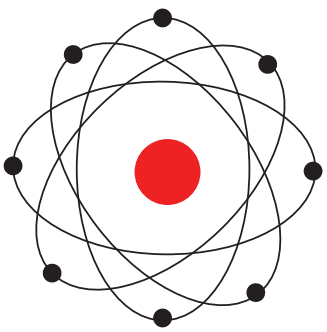


FIGURE 3.32 Rutherford's atom. Matter is mostly empty space.

The problem is that according to classical models the Rutherford model cannot lead to stable atoms. Rutherford's electrons are undergoing a centripetal acceleration and should radiate electromagnetic waves of the same frequency, so-called *bremsstrahlung* or "braking" radiation, leading to an electron "falling on a nucleus" in about 10^{-12} s! In the real world we have stable atoms, and atoms emit certain discrete characteristic frequencies of electromagnetic radiation. The Rutherford model is unable to explain these phenomena.

It was the analysis of light emitted or absorbed by atoms and molecules that led the way to better and better atom models. Cold, dilute gases absorb light at characteristic and discrete wavelengths (absorption spectra), and hot gases emit light at discrete wavelengths while continuous light spectra result when hot solids, liquids, very dense gases, or blackbodies (see Figure 3.33) emit light at all wavelengths (emission spectra).

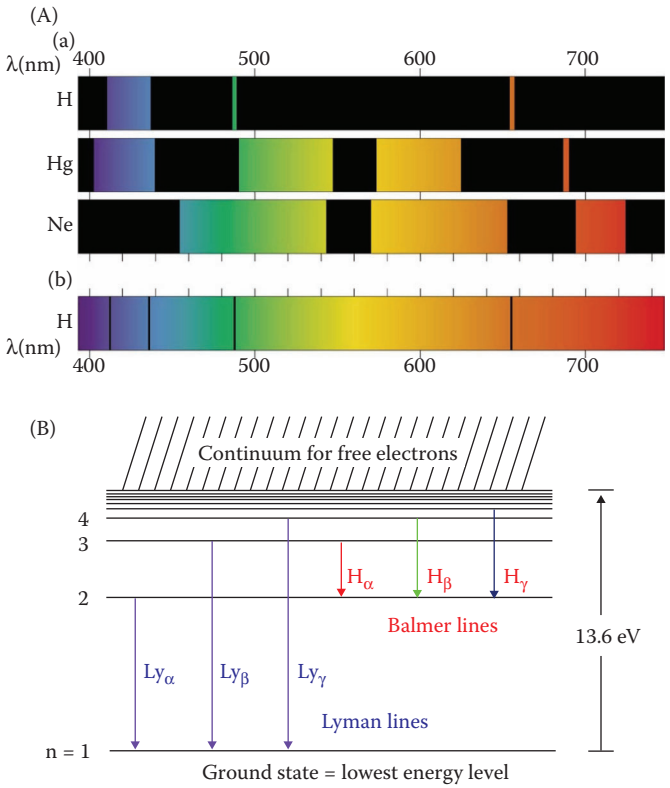


FIGURE 3.33 (Aa) Emission lines for H, Hg, and Ne; (Ab) Spectrum of sunlight with the hydrogen lines adsorbed out. (B) The hydrogen atom emission spectra with Balmer and Lyman lines for hydrogen.

When the light emitted from a hot gas is analyzed with a spectrometer, a series of discrete bright lines is observed. Each line has a different wavelength and color. This series of lines is the emission spectrum (Figure 3.33A). The absorption spectrum consists of a series of dark lines superimposed on the otherwise continuous spectrum. The dark lines of the absorption spectrum coincide with the bright lines of the emission spectrum. The continuous spectrum emitted by the sun passes through the cooler gases of the sun's atmosphere. The various absorption lines can be used to identify elements in the solar atmosphere, and this led, for example, to the discovery of helium.

Atomic hydrogen (one electron and one proton) in a gas discharge tube emits strongly at visible wavelengths H_α, H_β, and H_γ. More lines are found in the ultraviolet region, and the lines get closer and closer until a limit is reached (Figure 3.33B). These same lines are also seen in stellar spectra from absorption in the outer layers of the stellar gas.

There are four bright lines in the hydrogen emission spectrum, and in 1885, a Swiss teacher, Johann Balmer (1825–1898), guessed the following formula for the wavelength of these four lines:

$$\lambda = 364.5 \times 10^{-7} \frac{n^2}{n^2 - 4} (m) \qquad (3.80)$$

where n = 3, 4, 5, and 6, which are now called the *Balmer series*. Equation 3.80 may be rewritten as the so-called Balmer formula as:

$$\frac{1}{\lambda} = R_H \left(\frac{1}{n^2} - \frac{1}{k^2} \right) \qquad (3.81)$$

with positive n and k integers and $n > k$ and R_H the Rydberg constant for hydrogen with a measured value of $1.096776 \times 10^7\ m^{-1}$. With n = 1, we obtain the Lyman series in the UV; visible light is emitted in the Balmer series (n = 2), and with n ≥ 3 the infrared series are obtained (Paschen with n = 3, Brackett with n = 4, and Pfund with n = 5). The discrete emissions in Figure 3.33B suggest discrete energy levels, and Balmer's formula suggests that the allowed energies are given by (R_H/n^2) (in case of hydrogen) (Figure 3.34).

The "Great Dane," Niels Bohr, explained the above results for the H emission spectra by introducing four quantum postulates in a model halfway

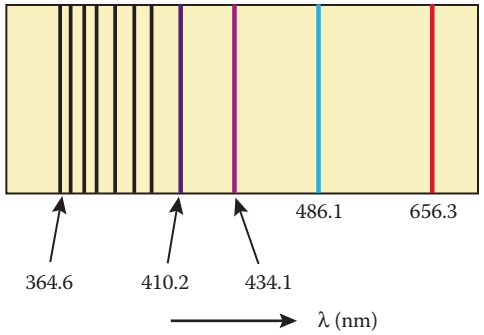


FIGURE 3.34 Some of the hydrogen emission lines.

between classical physics and quantum theory. He reasoned that if electrons orbit the nucleus in circles with radii restricted to certain values, then the energy also can only take on certain discrete values, i.e., if it is quantized and there is a lowest energy orbit, and the electron is not allowed to fall to a lower energy. The allowed states are called *stationary states*. When in these permitted orbits, contrary to classical theory, the electrons do not radiate (Postulate 1). Bohr also assumed that classical mechanics applies to electrons in those stationary states (Postulate 2). He recognized that there might be a link between stable orbits and the Planck's and Einstein's quantum relation between the quantized energy of a photon and its frequency, so he proposed that radiation absorption or emission corresponds to electrons moving from one stable orbit to another, i.e., $\Delta E = h\nu$ (where h is the Planck constant = 6.6×10^{-34} J·s) (Postulate 3). In classical physics, angular momentum (L) of an object in circular motion is defined as mass (m_e) multiplied by velocity (v), multiplied by the radius (r) of the orbit, i.e., $L = m_e v r$ (or $L = pr$) (Figure 3.35). Bohr argued that allowed orbits are determined by the quantization of that angular momentum, $L = nh/2\pi$.

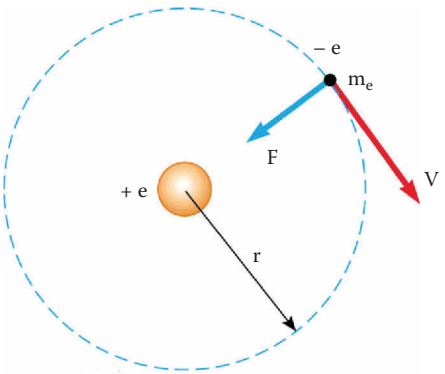


FIGURE 3.35 Momentum **p** for circular motion is $m_e vr$.

In this expression n is called the principal quantum number, and the ground state corresponds to n = 1. So here Bohr postulated that it was not the atom that determined the Planck constant h, but h that determined the properties of atoms (Postulate 4)!

Although this planetary kind of model has been shown to be mostly wrong, it makes for a very nice transition to full-fledged quantum mechanics. Bohr's theory, with slight modifications, is used, for example, to estimate the binding energies of dopant atoms in Chapter 4 and to explain the energy of excitons in Chapter 5.

Mathematically we can express Postulate 4 in vector notation as $|\mathbf{L}| = |\mathbf{p} \times \mathbf{r}| = m_e vr$, with v the tangential velocity and $\mathbf{p}$ the momentum of the electron (Figure 3.35). Bohr now combined his quantum model of the atom with Newton's classical model of planetary orbits to calculate the radius of the hydrogen atom. The attractive force between the electron and the proton in a hydrogen atom is caused by the Coulomb force and is given as:

$$F = \frac{e^2}{4\pi\varepsilon_0 r^2} \tag{3.82}$$

Based on Newton (Postulate 2), we may write:

$$F = \frac{m_e v^2}{r} = \frac{e^2}{4\pi\varepsilon_0 r^2} \tag{3.83}$$

Rutherford could not explain the spectral lines in and Figures 3.33 and 3.34, but he did set up the energy balance for a H atom correctly:

$$E_n \text{ (total energy)} = \text{KE (kinetic energy)} + \text{PE (potential energy)} \tag{3.84}$$

where the KE term depends on the velocity v and the PE term on the system (e.g., positional or electrostatic), or in the current case:

$$E = \frac{m_e v^2}{2} - \frac{e^2}{4\pi\varepsilon_0 r} \tag{3.85}$$

The radial acceleration is $a = v^2/r$, and with Newton's Law (F = ma) this yields $F = e^2/(4\pi\varepsilon_0 r^2) = mv^2/r$, which, solving for r, results in:

$$r = e^2/4\pi\varepsilon_0 m_e v^2 \tag{3.86}$$

And substituting Equation 3.86 in Equation 3.85 results in:

$$E = \frac{m_e v^2}{2} - m_e v^2 = -\frac{m_e v^2}{2} < 0 \tag{3.87}$$

meaning that, because E < 0, the motion of the electron is not free: it is bound by the attractive force of the nucleus as illustrated in Figure 3.36. To free the electron (and ionize the atom), the electron must receive an amount of energy, called

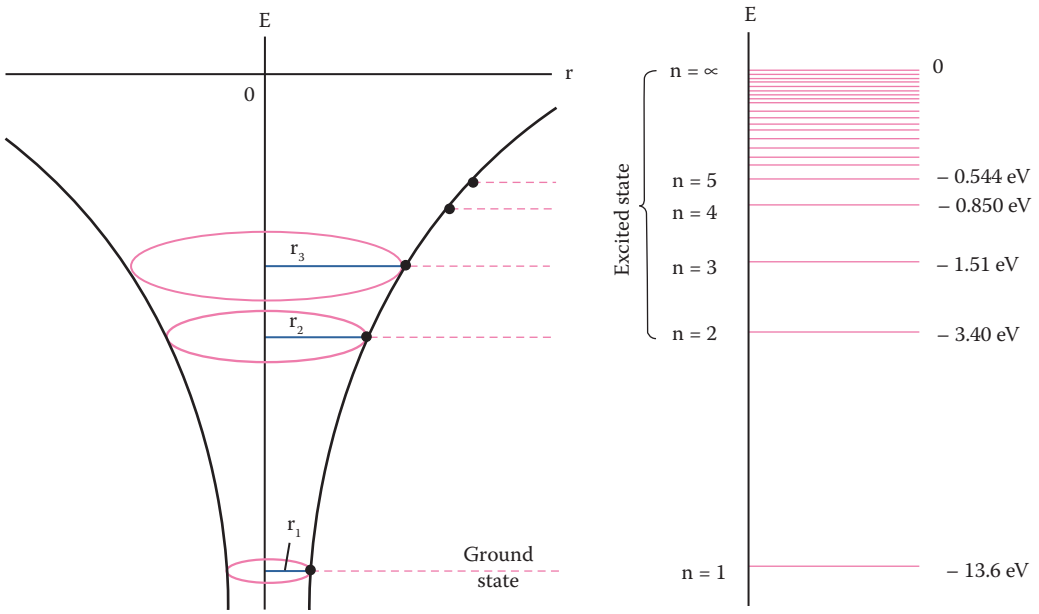


FIGURE 3.36 The electron is bound by the attractive force of the nucleus $E < 0$.

ionization energy, that will bring its total energy up to zero. Although this concept of a negative energy is somewhat counterintuitive, it does make sense; if we say that an electron has zero energy when far removed from the nucleus, then the electrons that are attached to an atom have a negative amount of energy.

The Quantum Numbers

First or Principal Quantum Number n for the Level of Energy E_n

Now we recall Bohr's fourth postulate about the quantization of angular momentum, $L_n = m_e v r_n = \frac{nh}{2\pi}$ or also $m_e v r_n = n\hbar$ with $\hbar = h/2\pi$. In the latter expression, n = 1, 2, 3, 4… and is called the principal quantum number; the ground state corresponds to n = 1. Thus, L_n is not only conserved but also constrained to discrete values by the quantum number *n*. This quantization of angular momentum is a crucial result and can be used in determining the Bohr orbit radii and Bohr energies. From the expression for *r* (Equation 3.86), we derive:

$$r_n = \frac{e^2}{4\pi\varepsilon_0 m_e (n\hbar/m_e r_n)^2} = \frac{e^2 r_n^2 m_e}{4\pi\varepsilon_0 n^2\hbar^2} \qquad (3.88)$$

Solving this expression for r_n results in:

$$r_n = \frac{n^2\hbar^2 4\pi\varepsilon_0}{m_e e^2} = n^2 a_0 \qquad (3.89)$$

where a_0 is the Bohr radius, and permitted orbits have radii $r_1 = a_0$, $r_2 = 4a_0$, $r_3 = 9a_0$ … for n = 1, 2, 3, … as shown in Figure 3.37.

The Bohr radius, $a_0 = \frac{\hbar^2 4\pi\varepsilon_0}{m_e e^2}$, has a value of 0.53×10^{-10} m.

We are now also in a position to calculate Bohr orbit speeds v_n:

$$v_n = \frac{n\hbar}{m r_n} = \frac{\hbar}{n m a_0} = \frac{e^2}{4\pi n\hbar\varepsilon_0} \qquad (3.90)$$

With the latter two expressions, we are able to calculate the total energy of the electron, E_n, associated

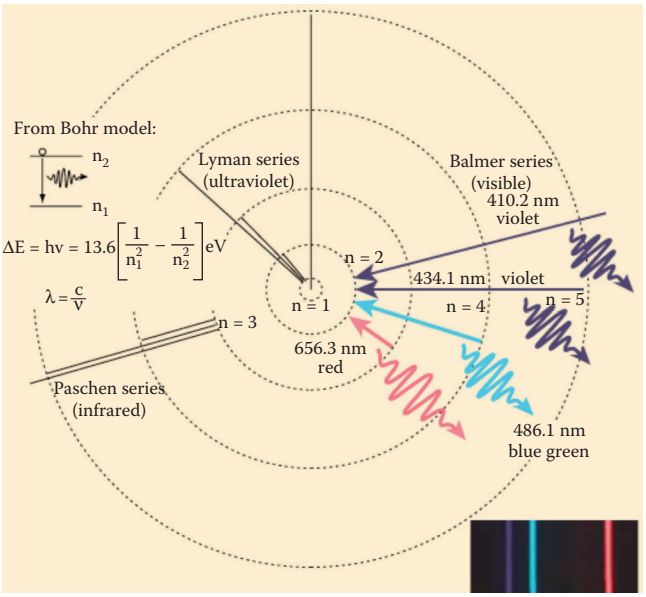


FIGURE 3.37 The Bohr atom with permitted radii. This Bohr model picture of the orbits has some usefulness for visualization, as long as it is realized that the "orbits" and the "orbit radius" just represent the most probable values of a considerable range of values.

with any integer value for *n*; by substituting the last two expressions in Equation 3.85, we obtain:

$$\begin{aligned} E_n &= \frac{m_e v^2}{2} - \frac{e^2}{4\pi\varepsilon_0 r_n} \\ &= \frac{1}{2} m_e \left[\frac{e^2}{n\hbar 4\pi\varepsilon_0}\right]^2 - \frac{e^2}{4\pi\varepsilon_0 n^2 a_0} \\ E_n &= \frac{m_e e^4}{32\hbar^2\pi^2\varepsilon_0^2 n^2} - \frac{m_e e^4}{16\hbar^2\pi^2\varepsilon_0^2 n^2} \\ &= -\frac{m_e e^4}{32\hbar^2\pi^2\varepsilon_0^2 n^2} = -\frac{E_0}{n^2} \end{aligned} \qquad (3.91)$$

and $E_0 = \frac{e^2}{8\pi\varepsilon_0 a_0} = \frac{me^4}{2(\hbar 4\pi\varepsilon_0)^2}$ with a value of 13.6 eV = 21.8×10^{-19} J, sometimes called the Rydberg energy. The 13.6-eV energy value for an electron in a hydrogen atom (with n = 1) is the energy required to remove an electron from that atom (see also Figure 3.33). The possible energy levels of the hydrogen atom are labeled by the values of the quantum number *n*. The lowest energy level occurs for n = 1; this is the most negative energy level and the ground state. As *n* progressively increases, the energy increases

(becomes less negative) for the excited states of the hydrogen atoms, as is clear from Figure 3.36.

Using the expression $E_n = -\frac{E_0}{n^2}$, we can now invoke Bohr's second postulate, which says that the photon energy is the difference in E_n values, and we derive Balmer's formula:

$$E_n = h\nu = \frac{hc}{\lambda} = -E_0\left[\frac{1}{k^2} - \frac{1}{n^2}\right]$$

$$\frac{1}{\lambda} = \frac{E_0}{hc}\left[\frac{1}{n^2} - \frac{1}{k^2}\right] = \frac{e^2}{8\pi\varepsilon_0 a_0 hc}\left[\frac{1}{n^2} - \frac{1}{k^2}\right] \qquad (3.92)$$

$$= R_H\left[\frac{1}{n^2} - \frac{1}{k^2}\right]$$

It was considered a big success for the Bohr model that it accounted for Balmer and other series and that the calculated R_H agreed almost exactly (to within 0.1%) with the phenomenological value for the Rydberg constant for hydrogen, i.e., $1.096776 \times 10^7\ m^{-1}$. The model also gives an expression for the radius of the atom that increases with n^2, the orbit speed that decreases with $1/n$, and predicts the energy levels of a hydrogen atom that increase with $-1/n^2$, and it can be extended to "hydrogen-like" atoms.

Second Quantum Number l for Orbital Angular Momentum

By 1914, Bohr had combined the quantum models of Planck and Einstein with the experimental work of Rutherford to provide a quantum model of the hydrogen atom, which fully explained the bright line spectra of hydrogen. To explain the spectra of more complicated atoms, other quantum numbers in addition to the principal quantum number n, defined by Bohr, had to be introduced. More detailed analysis of light spectra again led the way. Arnold Sommerfeld's (1868–1951) first major contribution was to extend Bohr's model to quantize all types of motion. He conceived elliptical atomic orbits by analogy to Johannes Kepler's elliptical planetary orbits and described mathematically orbits with different elliptical shapes but the same value of the principal quantum number n. This gave a number of different stationary states, some with slightly smaller and some with slightly larger energies—and hence multiple spectral lines—just as observed. The orbital angular momentum for an atomic electron can be visualized in terms of a vector model where the angular momentum vector is seen as precessing about a direction in space. Whereas the angular momentum vector has the magnitude shown in Figure 3.38, only a maximum of l units of $\hbar$ can be measured along a given direction, where l is the orbital quantum number. One of Sommerfeld's other important contributions was to link quantum theory to relativity. If one calculates the speed of an electron moving around the nucleus of an atom a value of the order of 1000 km/s is found. This is small compared with the velocity of light but large enough to have to use relativistic mechanics (see Chapter 5). Using this approach Sommerfeld was able to work out the fine structure of the hydrogen spectrum—a result that was regarded as a triumph both for relativity and quantum theory.

Third Quantum Number m, the Magnetic Number

In the 1890s, Pieter Zeeman (1865–1943) had noticed that if a magnetic field **B** was applied to a hot gas, the emission lines were further split into yet a finer structure. Sommerfeld, in 1916, showed that this was because of the direction (i.e., the orientation) of the orbiting of the electron with respect to the magnetic field. Sommerfeld was able to account for this orientation effect with the addition of a magnetic quantum number, m, which was an integer (Figure 3.39). Although called a "vector," the

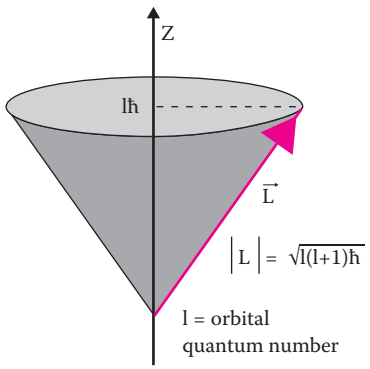


FIGURE 3.38 The Bohr-Sommerfeld atom. Sommerfeld introduced a second quantum number, *l* for elliptical orbitals. Spin-up and spin-down. Pauli's fourth quantum number: the electron spin *s* (see text later this chapter and Figure 3.39).

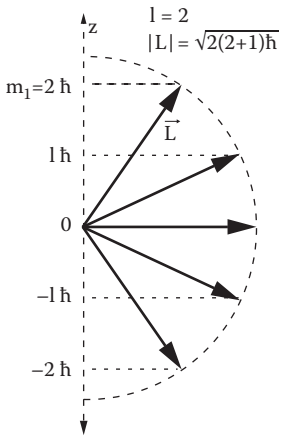


FIGURE 3.39 The quantum number m describes how the direction of the angular momentum is quantized with respect to the direction of the magnetic field.

orbital angular momentum in quantum mechanics is a special kind of vector because its projection along a direction in space is quantized to values one unit of angular momentum ($\hbar$) apart. In Figure 3.39 we show that the possible values that the "magnetic quantum number," m_l (for l = 2), can take are the values m = –2, –1, 0, +1, +2.

Fourth Quantum Number s for Electron Spin

More detailed measurements of the effects of a magnetic field on spectral line splitting showed yet further fine double splitting of the spectral lines. Wolfgang Pauli (1900–1958) explained this a little later, in 1921, by introducing yet another quantum number. While still a student, he proposed a fourth quantum number that he took to represent the electron spinning, not just in its orbit, but around its own axis. Because there are only two directions of spin, clockwise or spin-up and counterclockwise or spin-down, there are only two values for Pauli's electron spin quantum number. No two electrons in an atom can have identical quantum numbers. This is an example of a general principle that applies not only to electrons but also to other particles of half-integer spin (fermions). It does not apply to particles of integer spin (bosons). The two lowest-energy electron shells have an almost identical shape. Of the two, one shell is occupied or "filled" first with an electron that has an intrinsic magnetic direction that is opposite to the intrinsic magnetic field caused by the nucleus. The next shell has an electron with the opposite magnetic direction. The Dutch-American physicists Samuel Goudsmit and George Uhlenbeck discovered the intrinsic "spin" magnetism of the electron in the 1920s. It had been discovered years earlier that in a magnetic field, a beam of electrons splits into two beams (Stern-Gerlach experiment). This fourth quantum number explains this phenomenon. It is believed to be caused by some internal circulation of the electron matter, in addition to its wave flow around the equator of the shell. The wave flow around the equator of the atom also produces atomic orbital magnetic effects. Some shells have no net orbital circulation, which is explained as the result of two equal and opposite counter-rotating orbital waves. The magnetism of the nucleus itself is the result of the fundamental internal spin of the proton.

Surprisingly, it turned out that the fourth quantum number, spin, *s*, had only half of the usual quantization value of $h/2\pi$, i.e., $s = \pm 1/2\ h/2\pi$. Each quantum state, characterized by *n*, *l*, and *m*, is restricted to one electron of s = +1/2 and one electron of s = –1/2. Here classical analogies break down completely—a spin of a half implies the electron has to turn round twice to get back to where it started!

The implications of Pauli's exclusion principle are extremely profound. It is the restrictions imposed by Pauli's exclusion principle—i.e., the fact no two electrons can be in the same quantum state—that prevents all the electrons from piling up into the lowest (n = 1) energy state, and hence stops all matter from collapsing! Pauli's exclusion principle also implies that there is some sort of connectivity between the electron states in an atom: one electron must "know" which states all the other electrons are occupying to choose its own state! It is part of one of our most basic observations of nature: particles of half-integer spin must have antisymmetric wave functions, and particles of integer spin must have symmetric wave functions.

In Table 3.5 the different quantum numbers with their properties are summarized. The quantum numbers associated with the atomic electrons along with Pauli's exclusion principle provide insight into the building up of atomic structures and the periodic properties observed. For a given principal number *n*, there are $2n^2$ different possible states.

Table 3.5 Quantum Numbers and Their Properties

The principal quantum number:	
n	The principal quantum number. Quantization of angular momentum: must be a positive integer (1, 2, 3, 4 … etc.).
The angular momentum quantum number:	
l	Related to the ellipticity of the orbit: must again be an integer but for a particular orbit can be no bigger than n (l = 0, 1, 2, 3 … n – 1).
The magnetic quantum number:	
m	Quantization of the orientation of the orbit with respect to a magnetic field: can be a positive or negative integer (m = –l, –l + 1, 0, 1, 2, … l) but must be no larger than $-l \le m_l \le 1$.
The spin quantum number:	
s	The electron spin quantum number: must be +1/2 or –1/2.

n	Possible l	Possible m	Possible s	Spectroscopic Notation	Total States	Shell/Maximum Number of Electrons
1	0	0	±1/2	1s	2	K or 1st/2
2	0 1	0 –1,0,+1	±1/2 ±1/2	2s 2p	2 6	L or 2nd/8
3	0 1 2	0 –1, 0, +1 –2, –1, 0, +1, +2	±1/2 ±1/2 ±1/2	3s 3p 3d	2 6 10	M or 3rd/10
4	0 1 2 3	0 –1, 0, +1 –2, –1, 0, +1, +2 –3, –2, –1, 0, +1, +2, +3	±1/2 ±1/2 ±1/2 ±1/2	4s 4p 4d 4f	2 6 10 14	N or 4th/32

The Periodic Table of Dmitri Mendeleev (1834–1907)

Most of the qualities of an atom are derived from the structure of its electron cloud with the nucleus in the hinterland. This includes most chemical, material, optical, and electronic properties. Not only did the quantum model of the atom proposed by Bohr, Sommerfeld, and Pauli explain atomic spectra, it also explained the periodic table itself (inside front cover plate)! Dmitri Mendeleev (1834–1907) had devised the periodic table by grouping together elements with similar chemical properties.

The order of filling of electron energy states in an atom is dictated by energy, with the lowest available state consistent with Pauli's principle being the next to be filled. The labeling of the levels follows the scheme of the spectroscopic notation. For each value of principal quantum number, one refers to a different shell: n = 1 for the *K* shell, n = 2 for the *L* shell, n = 3 for the *M* shell, and n = 4 for the *N* shell. Different values of *l* correspond to different subshells. For example, for n = 2 (*K* shell) we have two subshells, namely, 2*s* and 2*p*. For historical reasons, the orbital quantum numbers *l* were given names associated with their appearance in spectroscopic emission and absorption patterns: l = 0 is given the letter *s* (for sharp), l = 1 is given the letter *p* (for principal), l = 2 is given the letter *d* (for diffuse), and l = 3 is given the letter *f* (for fundamental). Therefore, we can write the electronic occupation of an atom's shell as nl^e, where *n* is the principal quantum number, *l* is the appropriate letter, and *e* is the number of electrons in the orbit. So, for example, nitrogen (N) with seven electrons has the configuration: $1s^2 2s^2 2p^3$. In chemistry, this is called the Aufbau (build-up) principle, from the German word for *structure*.

As the periodic table of the elements is built up by adding the necessary electrons to match the atomic number, the electrons will take the lowest energy consistent with Pauli's exclusion principle. The maximum population of each shell is determined by the quantum numbers, and the diagram in Figure 3.40 is a convenient way to illustrate the order of filling of the electron energy states. For a single electron, the energy is determined by the principal quantum number *n*, and that quantum

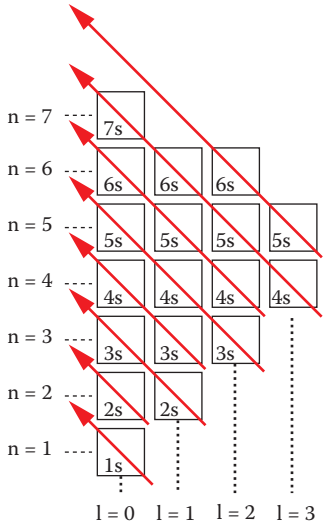


FIGURE 3.40 The periodic table of the elements is built up by adding the necessary electrons to match the atomic number.

number is used to indicate the "shell" in which the electrons reside. For a given shell in multielectron atoms, those electrons with lower orbital quantum number l will be lower in energy because of greater penetration of the shielding cloud of electrons in inner shells. These energy levels are specified by the principal and orbital quantum numbers using the spectroscopic notation. When you reach the 4*s* level, the dependence on orbital quantum number is so large that the 4*s* is lower than the 3*d*. Although there are minor exceptions, the level crossing follows the scheme indicated in the diagram, with the arrows indicating the points at which one moves to the next shell rather than proceeding to a higher orbital quantum number in the same shell.

The quantum scheme gives a firm scientific basis for Mendeleev's grouping: the chemical properties of the elements are defined by how nearly full or nearly empty a shell is. For example, full shells are associated with chemical stability (e.g., helium, neon, argon). Shells with a single electron or with one electron short of a filled shell are associated with chemical activity (e.g., sodium, potassium, chlorine, bromine). Most metals are formed from atoms with partially filled atomic orbitals, e.g., Na and Cu, which have the electronic structure:

Na $1s^2\,2s^2\,2p^6\,3s^1$
Cu $1s^2\,2s^2\,2p^6\,3s^2\,3d^{10}\,4s^1$

In the simplest picture, a metal has core electrons that are bound to the nuclei and valence electrons that can move through the metal. Insulators are formed from atoms with closed (totally filled) shells, e.g., inert gases:

He $1s^2$
Ne $1s^2\,2s^2\,2p^6$

or they form closed shells by covalent bonding as in the case of diamond.

With principal quantum number n = 4, the 4*f* orbitals make their capricious appearance. Because electrons in these orbitals are less strongly held by the nucleus, it is easier to excite them, and they can exhibit a myriad of distinct energy states. This gives rise to the unusual optical and magnetic properties of the rare-earth elements.

Most atoms with odd atomic numbers (1, 3, 5…) have a very slight overall atomic magnetism because of the one electron spin (and some orbital magnetism in some elements), whereas most even atomic number (2, 4, 6…) atoms have no net electron spin magnetism, and thus approximately zero resulting atomic magnetism.

Bohr's Correspondence Principle

Classical physics works for large systems; it is only at the atomic level that it fails. In 1923, Bohr proposed that any satisfactory quantum theory, then still being sought, should be in agreement with classical physics when the energy differences between quantized levels are very small or the quantum numbers are very large. This is the so-called Bohr correspondence principle. Let us consider the two main differences between the quantum theory and classical physics. The first difference is that whereas classical theory deals with continuously varying quantities, quantum theory deals with discontinuous or indivisible processes (e.g., the unit of energy packed in a quantum). The second difference is that whereas classical theory completely determines the relationship between variables at an earlier time and those at a later time, quantum laws determine only probabilities of future events in terms of given conditions in the past. The Bohr correspondence principle states that the laws of quantum physics must be so chosen that in the classical limit, where many quanta

are involved (e.g., n is a large integer in E = nh), the quantum laws lead to the classical equations as an average. This requirement combined with indivisibility and incomplete determinism defines the quantum theory in an almost unique manner. In this chapter we will on a few occasions show how classical theory can be derived from an extension of quantum mechanics. An example of this principle that we may appreciate already is that of Newtonian mechanics as a special case of relativistic mechanics when the speed is much less than the speed of light ($v \ll c$) (see Chapter 5, Equation 5.180). Similarly, Schrödinger's equation agrees with Bohr's ideas in general, except that the electron wave function is not at a given radius but spread out over a range of radii.

Summary of Bohr's Model

Although Bohr's model, in which r (space), energy, and momentum are quantized, was a major step toward understanding the quantum theory of the atom, it is not in fact a correct description of the nature of electron orbits. Some of the shortcomings of the model are:

1. It fails to provide any understanding of why certain spectral lines are brighter than others.
2. There is no mechanism for the calculation of electron transition probabilities.
3. Bohr's model treats the electron as if it were a miniature planet, with definite radius and momentum. This is in direct violation of the uncertainty principle (see below), which dictates that position and momentum cannot be simultaneously determined.

The precise details of spectra and charge distribution must be left to quantum mechanical calculations. Wave functions derived from Schrödinger's equation, the model that corrects all of Bohr's errors, are determined by the values of the above reviewed quantum numbers. For each energy level E_n, there is more than one distinct state with the same energy but different quantum numbers; this is called *degeneracy*. Degeneracy has no counterpart in Bohr's model. The solution of Schrödinger's equation, introduced below, for the case of the hydrogen atom is achieved by using spherical polar coordinates and by separating the variables so that the wave function is represented by the product of three terms. The separation leads to three equations for the three spatial variables, and their solutions give rise to three quantum numbers associated with the hydrogen energy levels. The solution to these equations can exist only when a few constants, which arise in the solution, are restricted to integer values. This gives the same hydrogen atom quantum numbers: n, l, and m as introduced. The fourth quantum number is, of course, again the electron spin s = 1/2, an intrinsic property of electrons.

de Broglie Matter Waves

Matter Waves

To launch Schrödinger's equation, we revisit de Broglie's concept of "matter waves." Louis-Victor, seventh duc de Broglie (1892–1987), discovered that the secret of Planck's and Einstein's quanta could be found in a general law of nature, i.e., the dual character of waves and particles.

Einstein's special relativity* (Chapter 5) allows one to calculate the momentum, **p**, of a photon starting from:

$$E^2 = (\mathbf{p}c)^2 + (m_0c^2)^2 \tag{3.93}$$

The rest mass, m_0, is zero for a photon, so one obtains for its momentum:

$$E^2 = (\mathbf{p}c)^2 + 0 \text{ or } E = \mathbf{p}c \tag{3.94}$$

And also, from Einstein and Planck, E = hν (the photoelectric effect), or:

$$\mathbf{p} = \frac{E}{c} = \frac{h\nu}{c} = \frac{h}{\lambda} \tag{3.95}$$

de Broglie, while studying for his PhD in Paris in 1924, postulated that Equation 3.95 also applied to a moving particle such as an electron, in which case λ is the wavelength of the wave associated with the moving particle, i.e., a "matter wave." Einstein commented about this fantastic insight: "de Broglie has lifted the great veil." From Equation 3.95, $\mathbf{p}\lambda = h$,

* For p = 0 Equation 3.93 leads to the familiar $E = m_0c^2$. This amounts to 511 keV for an electron and 930 GeV for a proton. Where atomic physics deals in eV to keV, nuclear physics deals in keV to MeV and particle physics involves GeV to TeV.

or momentum multiplied by wavelength, gives the Planck constant; the smaller the wavelength, the bigger the momentum of the particle! Thus, electrons, with their small mass and correspondingly small momentum, are very "wavy" particles. The de Broglie wavelength of a particle is then given as:

$$\lambda = \frac{h}{mv} \tag{3.96}$$

where v is its velocity. This simple equation proves to be one of the most useful, and famous, equations in quantum mechanics. It accounts for both waves (characterized by wavelength and frequency) and particles (characterized by position, mass, and velocity), incorporating momentum (particle aspect) and the wavelength (wave aspect).

The critical reader could very legitimately ask here: "But how can photons have a momentum if they do not have any mass?" The answer is that photons do not have a mass, but they do have energy, and as Einstein famously proved, mass and energy are the same thing. For photons, the wavelength and frequency, from Equation 3.95, are related because $\lambda\nu = c$ is fixed, but for matter particles $\lambda\nu$ is not a constant because they do not travel at speed c. For nonzero rest mass particles, such as electrons or protons, neither $\lambda\nu = c$ nor $E = \mathbf{p}c$ applies.

We almost have come full circle here; Einstein gave what we had come to think of as a wave (light) a particle character, and de Broglie gave what we thought of as a particle (electrons) a wave character. Radiation has wave character and particle character, and matter has particle and wave character, or at the nanoscale, nature presents itself with a wave-particle duality. This duality for material things is only significant at the very small scale. With its mass m_e of 9.11×10^{-31} kg and a velocity v of 3×10^6 m/s (i.e., c/100), an electron has a wavelength of 2.5×10^{-10} m. Larger objects, such as a car (1000 kg), moving at highway speeds (60 mph ≈ 100 km/h) have a de Broglie wavelength as well, but there is no measurement capable of resolving them; the oscillations cannot be seen. In other words, the car does not behave like a wave, with a $\lambda = \frac{h}{p} = \frac{h}{mv} = \frac{6.6 \times 10^{-34}\,\text{J·s}}{1000\ \text{kg} \times 10^5\ \text{m/h} \times 1\,\text{h}/36{,}000\,\text{s}} = 2.4 \times 10^{-38}$ m. A man (60 kg) running at 20 km/h (5 m/s) has a wavelength of 2.2×10^{-36} m (for some other matter wave examples see Table 3.6).

Wave Packets

Because we associate a de Broglie wave with a moving body, it is reasonable to expect that this wave moves at the same velocity as that of the particle—this turns out not to be the case. de Broglie hypothesized that the particle itself was not a wave but always had with it a pilot wave, or a wave that helps guide the particle through space and time. He also pointed out that in wave theory there is a difference between the speed of waves of some uniquely defined frequency and the speed of a localized pulse. Waves have both a phase velocity and a group velocity, and de Broglie proposed to associate the localized pulse, wave packet, or group velocity with the moving particle.

To understand *matter* waves and their link with wave packets or group velocity and phase velocity better, we reconsider here for a moment a sinusoidal wave function $\psi(x) = Ae^{ikx} = A\cos kx + iA\sin kx$

Table 3.6 de Broglie Wavelengths of Various Moving Objects

	Mass	Kinetic Energy (eV)	de Broglie Wavelength (m)
Electron	9.11×10^{-31}	1	1.23×10^{-9}
		100	1.23×10^{-10}
		10^4	1.23×10^{-11}
Neutron	1.67×10^{-27}	1.00×10^3	9.05×10^{-13}
		1.00×10^6	2.86×10^{-14}
		1.00×10^9	9.05×10^{-16}
Proton	1.67×10^{-27}	1.00	2.86×10^{-11}
		1.00×10^2	2.86×10^{-12}
		1.00×10^3	9.04×10^{-13}
Thermal neutrons (300 K)		2.50×10^{-2}	1.81×10^{-10}

that extends all the way from $x = -\infty$ to $x = +\infty$ with an amplitude A. To make this wave more localized or "lumpy," we add together two different stationary-state waves with slightly different wave vectors $\mathbf{k}$ (e.g., $\mathbf{k}_1$ and $\mathbf{k}_2$) and different amplitudes A (e.g., A_1 and A_2) and consider the result at one instance of time, e.g., at $t = 0$. For $t = 0$, the time factor of the superposed waves, $e^{-i\omega t}$ (e.g., ω_1 and ω_2), is equal for both superposed waves, i.e., $e^0 = 1$, so the wave function is represented by:

$$\begin{aligned}\psi(x) &= A_1 e^{ik_1 x} + A_2 e^{ik_2 x} \\ &= [A_1 \cos(k_1 x) + iA_1 \sin(k_1 x)] \\ &\quad + [A_2 \cos(k_2 x) + iA_2 \sin(k_2 x)] \\ &= [A_1 \cos(k_1 x) + iA_2 \cos(k_2 x)] \\ &\quad + i[A_1 \sin(k_1 x) + A_2 \sin(k_2 x)]\end{aligned} \tag{3.97}$$

This is represented in Figure 3.41a, where we graph the real parts of the individual waves for the case that $A_2 = -A_1$. The real part of the combined wave function is shown in Figure 3.41b. With two additional sinusoidal waves, we can choose to emphasize every other lump and suppress the in-between ones as illustrated. Going on like this, we can, in the extreme, superpose a very large number of waves with different $\mathbf{k}$ and A values to construct a wave with only one lump, as illustrated in Figure 3.41c. The latter wave pulse is called a *wave packet*. The particle represented by the lump in Figure 3.41c is more likely to be found in the region of the lump than in other regions; in other words, it has become more localized. However, the particle now does have a less definite momentum as we introduced a very large number of different waves. The above procedure of adding waves is not new to us: in Chapter 2 on Fourier transforms (Equations 2.15 and 2.16) we saw how a localized disturbance or pulse can be synthesized from a very large number of pure sinusoidal waves of different frequencies and amplitudes. Such a pulse, a wave function with a lump, is an entity with both particle and wave characters. It behaves like a particle in the sense that it is localized in space, but it also has periodicity, a property of a wave. The Fourier integral representing the synthesized pulse or wave packet is:

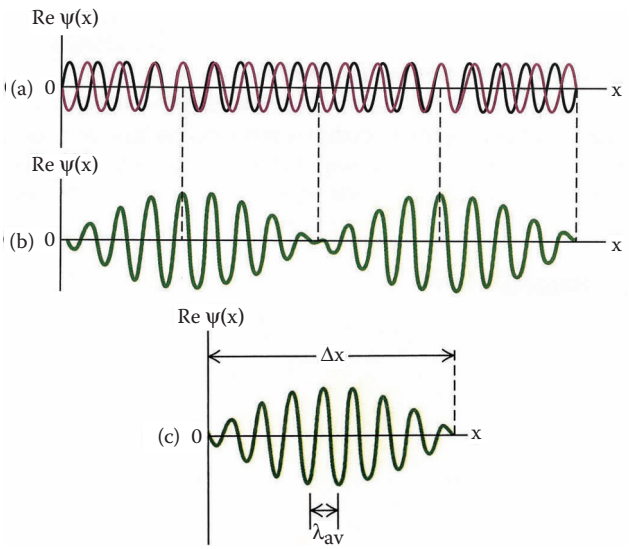


FIGURE 3.41 (a) Two sinusoidal waves with slightly different wave numbers at one instant of time ($A_2 = -A_1$) in Equation 3.97. (b) The superposition of these two waves has a wave number equal to the average of the two individual wave numbers. The amplitude varies, giving the total wave a more lumpy character. (c) Superimposing a large number of sinusoidal waves with different wave numbers and appropriate amplitudes can produce a wave pulse.

$$\psi(x) = \int_{-\infty}^{\infty} A(\mathbf{k}) e^{ikx} d\mathbf{k} \tag{3.98}$$

This Fourier integral superimposes wave functions with a continuous distribution of $\mathbf{k}$ values and amplitudes $A(\mathbf{k})$ that depend on $\mathbf{k}$. Below we will see how Heisenberg's uncertainty principle (HUP) follows directly from the existence of the de Broglie waves and this Fourier integral (see Figure 3.47 further below).

The synthesized pulse travels with a group velocity (V_g) that may be different from the characteristic phase velocity (V_p) of the individual waves. Let us consider the Fourier superposition of just two waves, $\Psi_1 = \sin[kx - \omega t]$ and $\Psi_2 = \sin[(k + \Delta k)x - (\omega + \Delta\omega)t]$, as represented in Figure 3.42a, where we can write:

$$\Psi_1 + \Psi_2 = 2\cos\left[\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x\right] \times \sin\left[\left(k + \frac{\Delta k}{2}\right)x - \left(\omega + \frac{\Delta\omega}{2}\right)t\right] \tag{3.99}$$

Modulation — Sine wave

The phase velocity is directly proportional to the angular frequency, but inversely proportional to the wave number, or the phase velocity v_p is given as:

$$v_p = \frac{\omega + \frac{\Delta\omega}{2}}{\mathbf{k} + \frac{\Delta\mathbf{k}}{2}} = \frac{\omega'}{\mathbf{k}'} \tag{3.100}$$

from the sine wave in Equation 3.99. For the group velocity v_g and thus the particle velocity:

$$v_g = \frac{\frac{\Delta\omega}{2}}{\frac{\Delta k}{2}} = \frac{\Delta\omega}{\Delta\mathbf{k}} = \frac{d\omega}{dk} \tag{3.101}$$

which is based on the modulation amplitude in Equation 3.99. This group velocity, the derivative of the angular frequency with respect to the wave vector, is the velocity at which the energy of the wave propagates. Group and phase velocities are illustrated in Figure 3.42b.

The group velocity is the derivative of the phase velocity, and it is often the case that the phase velocity is larger than the group velocity. For any wave that is not electromagnetic, the phase velocity will be larger than *c*, or the speed of light.

de Broglie was trying to find a velocity that was fit for all particles (not just photons), so he associated the group velocity v_g with any particle's velocity. He equated the following two expressions:

$$v_g = \frac{\partial\omega}{\partial\mathbf{k}} = \frac{\partial E}{\partial\mathbf{p}} \tag{3.102}$$

because $v_p = E/\mathbf{p}$ and $v_g = \partial E/\partial\mathbf{p}$ and (as we will see below) $E = \hbar\omega$ and $p = \hbar\mathbf{k}$, where $\hbar = h/2\pi$. In the case of light:

$$v_p = \frac{E}{\mathbf{p}} = \frac{mc^2}{mc} = c \text{ and } v_g = \frac{\partial E}{\partial p} = \frac{\partial pc}{\partial p} = c \tag{3.103}$$

We get the same result for phase and group velocity. For any other particle, from Equation 3.93 with $m_0 = (m \neq 0)$, we obtain:

$$v_g = \frac{\partial E}{\partial p} = \frac{pc^2}{\sqrt{(\mathbf{pc})^2 + \left(mc^2\right)^2}} = \frac{pc^2}{E} \tag{3.104}$$

with $m = m_0$, $v_g = c$ again.

de Broglie Vindicated

de Broglie's doctoral dissertation exam committee was distinctly unsure as what to do with the thesis of

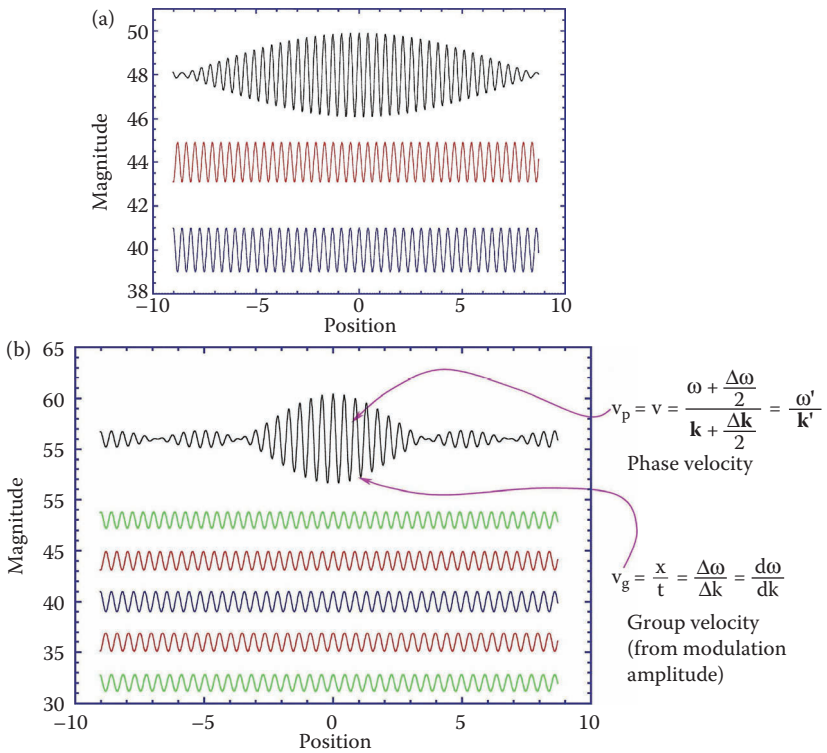


FIGURE 3.42 (a) Using Fourier's theorem to construct wave packets. (b) Phase velocity v_p from the sine wave and group velocity v_g from the modulation amplitude.

the prince. Luckily, Langevin, who was on the committee, asked Einstein's opinion on the work, and, as mentioned above, Einstein was duly impressed. Experimental evidence was soon to follow.

J.J. Thompson received the Nobel Prize in 1906 for his work demonstrating the particle nature of an electron. J.J. Thompson's son, G.P. Thompson, received the Nobel Prize in 1937 for experiments that tested Equation 3.96, demonstrating that electrons also exhibit wave-like properties. In his 1928 tests, G.P. Thompson and Reid at the University of Aberdeen observed interference patterns from electrons reflecting from a thin polycrystalline metal foil surface. Working independently, Clinton Davisson and Lester Germer, at Bell Laboratories, in 1927 found the same experimental evidence: a beam of electrons scattered from a single-crystal of nickel resulted in a diffraction pattern fitting Bragg's diffraction law (Equation 2.20 and Figure 3.43). This established the wave character of electrons, forming the basis of analytical techniques for determining the structures of molecules, solids, and surfaces, such as in low energy electron diffraction (LEED) and SEM described in Volume III, Chapter 6 on metrology. Whereas optical microscopes cannot resolve details smaller than the wavelength of visible light, with an SEM the wavelength depends on the electron's momentum in accordance with de Broglie's relation. In this case we can increase the resolution by simply accelerating the electron more.

Other particles, including protons, neutrons, and He atoms, were subsequently found to possess wave properties, including diffraction. In 1999, researchers from the University of Vienna demonstrated that the wave-particle duality even applied to molecules such as fullerene. One of the coauthors of this research, Julian Voss-Andreae, became an artist and has since created several sculptures symbolizing wave-particle duality in buckminsterfullerenes.[2]

Bohr's Model and de Broglie Matter Waves

de Broglie's idea of matter waves fits nicely with Bohr's model; the de Broglie wavelength associated with the electron is what causes the quantization that Bohr had assumed before matter waves were even known. The angular momentum of the electron is restricted to certain values because an integral number of electron wavelengths must fit into the circumference of the circular orbit. Mathematically we expressed Bohr's Postulate 4, in vector notation, as $|\mathbf{L}| = |\mathbf{p} \times \mathbf{r}| = m_e vr$, with v the tangential velocity and $\mathbf{p}$ the momentum of the electron. Bohr's model, as we saw above, assumed $\mathbf{L} = 2\pi r$, or:

$$\mathbf{L} = 2\pi r = n\lambda = \frac{nh}{\mathbf{p}} \text{ and } r = \frac{n\hbar}{p} (\text{with } \hbar = \frac{h}{2\pi}) \tag{3.105}$$

or

$$L = \mathbf{p}r = \frac{nh}{2\pi} = n\hbar$$

The latter is illustrated in Figure 3.44, where we show that only an integer number of wavelengths fits into a circular orbit or, i.e., $L = nh/2\pi$ (where L

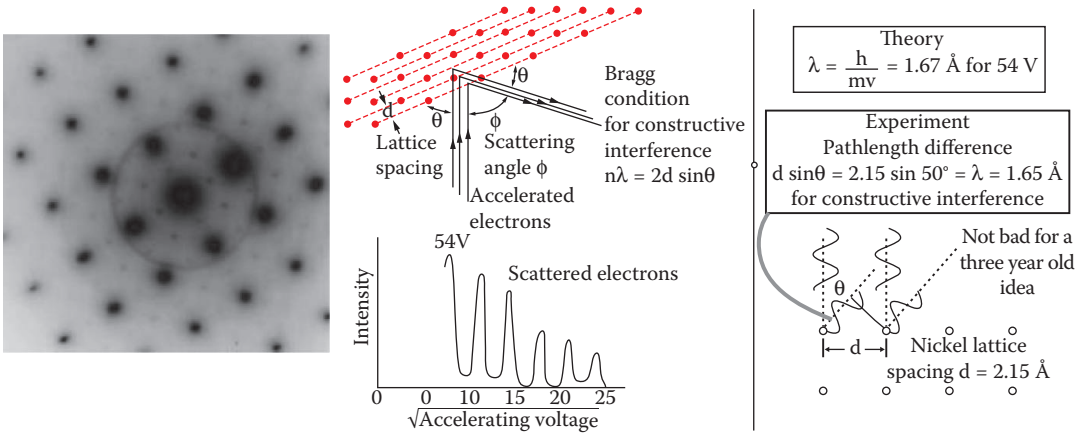


FIGURE 3.43 Davisson-Germer experiment (1927). A graph of the intensity of the scattered electron beam demonstrates that the angular position of the maxima depended on the accelerating voltage used to produce the electron beam. Using the de Broglie equation, the de Broglie wavelength of the electron could be calculated.

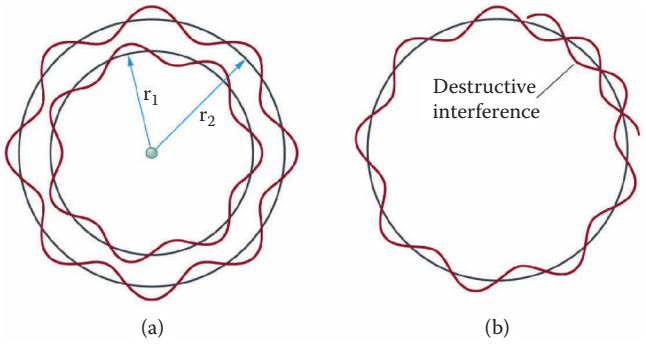


FIGURE 3.44 Bohr's atom model: only an integer number of wavelengths fits into a circular orbit, $L = nh/2\pi$ (where *L* is the length of an orbit and *h* is the Planck constant).

is the length of an orbit). Thus, a wave-mechanical picture leads naturally to the quantization of the angular momentum.

de Broglie's relations imply the quantization of orbits of electrons in atoms, $L = n\hbar$ with *n* an integer, but although this explains the root of quantization—the wave nature of matter—it does not provide a complete dynamic that would be able to describe the motion of particles that are waves at the same time. One needs a wave equation for that, and Maxwell's equations are inappropriate for this purpose because they predict that $\lambda\nu = c$, which is not valid for matter particles, where $\nu\lambda = v$. Interestingly, we will see next that Bohr's model violates Heisenberg's uncertainty principle: an electron may not move in a circle with exact radius *r*!

Heisenberg's Uncertainty Principle

The wave-particle duality introduced in the previous section forced physicists to reconsider their description of the position and momentum of very small particles, and it is at the core of Heisenberg's uncertainty principle (HUP). In Newtonian mechanics, we can always describe a particle in terms of its spatial coordinates and the three components of its velocity. However, in the nanoworld, Heisenberg's principle states that there are physical parameters in quantum physics whose values cannot be known accurately simultaneously. For example, the momentum, $\mathbf{p}_x$, and position, *x*, of an electron cannot be known simultaneously. In classical physics, knowing $\mathbf{p}_x$ would not be difficult because light is considered to have continuously varying energies and would hence cause minimal disturbance on an electron that is being observed. In quantum physics, light consists of photons that have discrete (quantized) energies, and a photon bouncing off an electron being observed gives that electron a kick, disturbing its momentum (Figure 3.45). To obtain an accurate measurement of the position of the electron, one must use a very short probing wavelength because a long wavelength would not have enough resolution to locate the electron. Thus, a short wavelength is expected to give good resolution or a small uncertainty in position. Nevertheless, a short wavelength

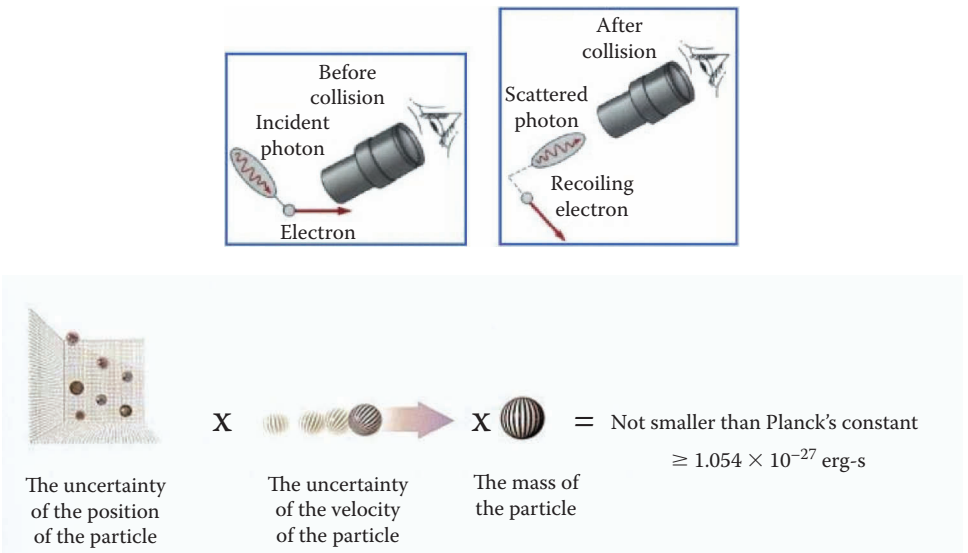


FIGURE 3.45 Looking is disturbing! Imagine trying to see an electron with a powerful optical microscope. At least one photon must scatter off the electron and enter the microscope, but in doing so it will transfer some of its momentum to the electron.

also means a large energy and hence a large momentum ($\mathbf{p}_x = \mathbf{E}/c$), which gives the observed electron too large a kick and too large an uncertainty about its momentum $\Delta\mathbf{p}_x$ (Compton effect).

A more quantitative analysis shows that the product of the two uncertainties, Δx and $\Delta\mathbf{p}_x$, is a constant:

$$\Delta\mathbf{p}_x\Delta x \geq \frac{h}{2\pi} = \hbar \qquad (3.106)$$

where h is again the Planck constant, and the constant "h-bar" has the approximate value of 10^{-34} J·s. There is an uncertainty relation as shown in Equation 3.106 for each coordinate and its corresponding momentum component. Equation 3.106 is graphically represented in Figure 3.46. Note that there is no restriction on the precision in simultaneously knowing/measuring the position along a given direction (x) and the momentum along another, perpendicular direction (y or z).

For a particle in circular motion, the equivalent Heisenberg expression to Equation 3.106 is $\Delta\mathbf{p}_r\Delta\mathbf{r} \geq \hbar$. Applying this to Bohr's model, in which an electron moves in a circle of *exact* radius r, we obtain $\Delta r = 0$ and $\Delta P_r = \infty$, or the model violates the uncertainty principle. We will see further below that despite this obvious problem, the energy level predictions of Bohr's model remain valid.

Similar ideas lead to the expression of uncertainty for other pairs of observables such as time and energy:

$$\Delta E\Delta t \geq \frac{h}{2\pi} = \hbar \qquad (3.107)$$

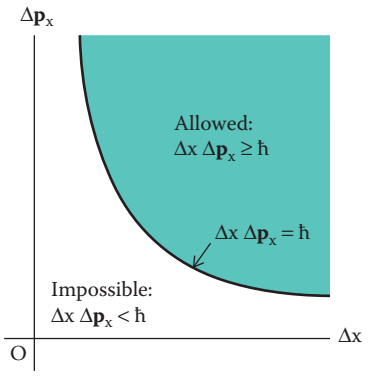


FIGURE 3.46 A graphic presentation of the Heisenberg principle from Equation 3.106 for position and momentum.

This says that if an energy state only lasts for a limited time, its energy will be uncertain. The uncertainty about the energy of a particle depends on the time interval Δt that the system remains in a given energy state. Importantly, this also means that conservation of energy can be violated if the time is short enough. From the uncertainty principles, it is possible that empty space locally does not have zero energy but may have sufficient ΔE for a very short time Δt to create particles and their antiparticles. This can be demonstrated through the Casimir effect (see below). Equation 3.107 is responsible for "lifetime broadening" of spectral lines. Short-lived excited states (small Δt) possess large uncertainty in the energy of the state (large ΔE). As a consequence, shorter laser pulses (e.g., femto- and attosecond lasers) have broader energy (therefore, wavelength) bandwidths.

The existence of a zero-point energy—vibrational energy cannot be zero even at T = 0 K (see below)—is also a consequence of Heisenberg's uncertainty principle. If the vibration would cease at T = 0 K, then the position and momentum would both be 0, violating the HUP.

One might object that perhaps light was a poor choice to measure the position and momentum of the electron and that some other method might avoid these uncertainties. No such luck: it turns out that this is the absolute best that can be achieved independently of the measuring method. Note that in the last two equations all references to light have dropped out; the result does not depend on λ, n, or c. The Heisenberg uncertainty principle follows solely from the wave-particle duality and has nothing to do with the unavoidable disturbance of the system by the measurement. Quantum mechanics tells us there are limits to measurement—not because of the limits of our instruments, but inherently.

Actually, the uncertainty principle is an inevitable consequence of de Broglie's relation and the Fourier integral representing the synthesized pulse or wave packet (Equation 3.98). In Figure 3.47 we illustrate qualitatively how $\psi(x)$ depends on $A(\mathbf{k})$. In Figure 3.47a, we show a sharp peak in $A(\mathbf{k})$ for a narrow range of wave numbers $\mathbf{k}$. The resulting real part of the wave pulse, shown in Figure 3.47b, is relatively broad: a narrow range of $\mathbf{k}$ means a narrow range of $\mathbf{p}_x = \hbar\mathbf{k}$ and thus a small $\Delta\mathbf{p}_x$, and the

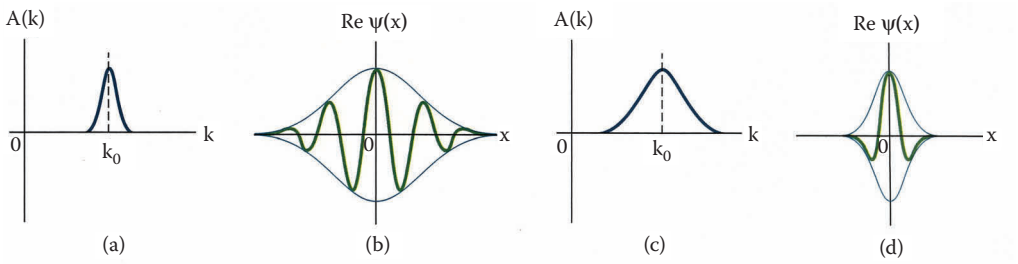


FIGURE 3.47 (a) A sharp peaked $A(\mathbf{k})$ function leads to a wave function $\psi(x)$ with a broad spatial extent (Δx) (b); (c) a broad peaked $A(\mathbf{k})$ function leads to a wave function $\psi(x)$ with a narrow spatial extent (Δx) (d).

result is a relatively large Δx. Broadening the $A(\mathbf{k})$ function in Figure 3.47c results in a more localized wave pulse with a smaller Δx, as clear from Figure 3.47d. In other words, this is Heisenberg's uncertainty principle, $\Delta \mathbf{p}_x \Delta x \geq \hbar$, in action.

Heisenberg's relations are of no practical importance in the macroworld, and in classical physics they can be ignored completely.

The Launchpad: Classical Mechanics Revisited

By 1924, the quantum concepts of Planck, Einstein, Bohr, and de Broglie were widely accepted, and between 1925 and 1927 three distinct, independent, and very different integrating theories of quantum theory were proposed: Dirac's Hamiltonian and quantum algebra representation; the matrix representation of Born, Heisenberg, and Jordan; and Schrödinger's wave equations. Much later, in the late 1940s, Feynman formulated his sum-over-histories approach.

Schrödinger, after attending a seminar on Einstein's and de Broglie's ideas that wave-like entities can behave like particles, and vice versa, thought that there must be a wave equation, $\Psi(x,t)$, to describe particles. Schrödinger almost completely dispensed with the concept of a particle and instead focused on the wave-like properties of matter. His picture of the atom has the electron standing waves vibrating in their orbitals much like the vibrations on a string—but in three dimensions instead of one. In Figure 3.48, a 2D representation of Schrödinger waves, like vibrations on a drum skin, is shown.

A concept that plays an important role in both classical and quantum theory is that of the Hamiltonian of a system. Consider an isolated system composed of one or more particles, and assume that the total energy of this system remains constant. The Hamiltonian of this system is merely its total energy:

$$\mathrm{H} = \mathrm{E_n}(\text{total energy}) = \mathrm{KE}\ (\text{kinetic energy, depends on v}) + \mathrm{PE}\ (\text{potential energy, depends on position}) \qquad (3.84)$$

with the kinetic energy arising from motion and the potential energy arising from the position in a force field, F. The Hamiltonian in quantum mechanics is an example of an operator, a mathematical object that tells us what operation to perform on the function that follows. We rewrite Equation 3.84 here as:

$$\mathrm{H} = \mathrm{E} = \mathrm{E_k} + \mathrm{V(x)} \qquad (3.108)$$

The linear momentum $\mathbf{p}$ equals mv so that $\mathrm{E_k} = \mathrm{mv}^2/2$ or $\mathbf{p}^2/2\mathrm{m}$. When working with the Hamiltonian, the kinetic energy of a particle is expressed as $\mathbf{p}^2/2m$, not as $mv^2/2$. Thus, Equation 3.84 may be rewritten as:

$$\mathrm{H} = \mathrm{E} = \frac{\mathbf{p}^2}{2\mathrm{m}} + \mathrm{V(x)} \qquad (3.109)$$

In case the potential is time varying, the last term in Equation 3.109 must be written out as $V(x,t)$. The total energy, E, in the absence of a potential energy

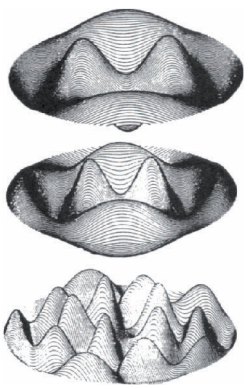


FIGURE 3.48 A two-dimensional representation of Schrödinger waves. Notice the nodes of the vibration.

equals E_k. Newton's second law relates potential energy to change in momentum. According to this law of mechanics, forces give rise to a change in the momentum of particles [(F = d(mv)/dt = dp/dt], and each force, *F*, has an associated potential energy *V*(*x*) (F = −dV/dx). The direction of the force is toward decreasing potential energy:

$$F = \frac{dp}{dt} = ma = m\frac{d^2x}{dt^2}$$
$$F = -\frac{dV(x)}{dx} \qquad (3.110)$$

or:

$$F = -\frac{dV(x)}{dx} = m\frac{d^2x}{dt^2}$$

i.e., given *V*(*x*) one can solve for *x*(*t*) or *v*(*x*,*t*). We illustrate the use of the above deterministic expressions with two important examples so we might better appreciate the analogies and the differences between classical and quantum physics.

Example 3.1: Free particle [no force, V(x) = 0] moving along *x* (1D problem):

$$E = E_k = \frac{mv^2}{2} = \frac{p^2}{2m} \qquad (3.111)$$

v = dx/dt, and:

$$v = \frac{dx}{dt} = \sqrt{\frac{2E_k}{m}} \qquad (3.112)$$

Therefore, we can formulate the differential equation:

$$dx = \sqrt{\frac{2E_k}{m}}dt \qquad (3.113)$$

Integration from x_0 to x_t and from t = 0 to t = t and using Equation 3.112 results in:

$$x_t - x_0 = \sqrt{\frac{2E_k}{m}}t = \sqrt{\frac{2p^2}{2m^2}}t = \frac{p}{m}t = vt \qquad (3.114)$$

With V = 0, knowing x_0 and *p*, we can predict x_t at *t*; in other words, we can predict the trajectory at all times later.

Example 3.2: Harmonic oscillators [F = −kx with *x* the displacement]. The pendulum, which is just a mass on a spring, is an example of a harmonic oscillator, but its physical description also encompasses many objects that oscillate, from tuning forks to oscillating bridges and oscillating skyscrapers (Figure 3.49).

Let us first calculate *V*(*x*) in Equation 3.109 for the given problem. The force F = −dV(x)/dx = −kx (Hooke's law), with *k* the force or spring constant, or:

$$dV(x) = kxdx \qquad (3.115)$$

Integrating, we obtain:

$$V - 0 = k\frac{1}{2}(x^2 - 0^2) \text{ since } V = 0 \text{ at } x = 0$$

or

$$V = \frac{kx^2}{2} \qquad (3.116)$$

The force F = ma = −kx, and we obtain the differential equation:

$$\frac{d^2x}{dt^2} = -\frac{k}{m}x \qquad (3.117)$$

with a solution:

$$x(t) = A\sin\left(\sqrt{\frac{k}{m}}t\right) \qquad (3.118)$$

where *A* is the maximum displacement (x_{max} = A). For a simple harmonic oscillator ω = 2πf = 2π/T and:

$$\omega = \sqrt{\frac{k}{m}} \quad \text{and} \quad T = 2\pi\sqrt{\frac{m}{k}} \qquad (3.119)$$

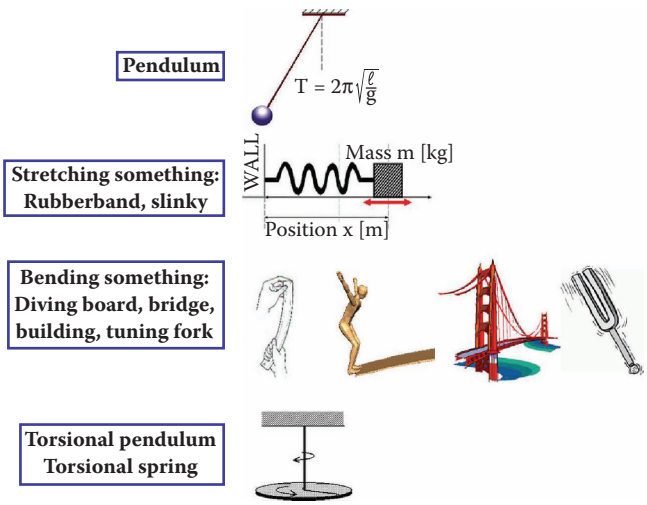


FIGURE 3.49 A wide variety of harmonic oscillator examples.

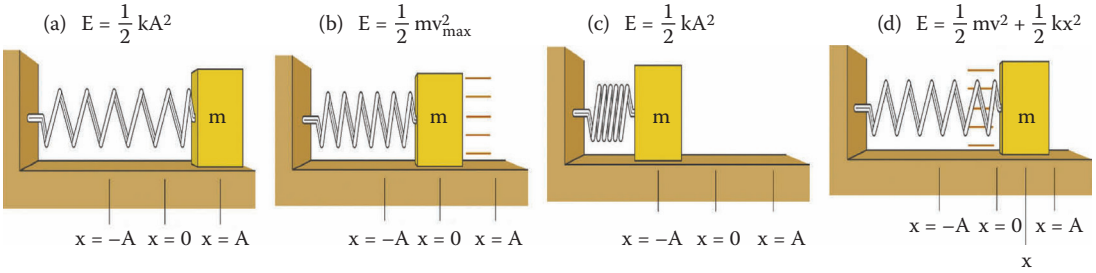


FIGURE 3.50 Spring/mass system with values for kinetic and potential energy as a function of the position of the mass on the *x*-axis. (a) x = A; (b) x = 0; (c) x = –A; and (d) x = x.

Observe that frequency only depends on characteristics of the system (*m*,*k*) and not on the amplitude *A*. The first and second derivatives of Equation 3.118 are given as:

$$v(t) = \frac{dx}{dt} = A\sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}t\right)$$

and

$$a(t) = \frac{d^2x}{dt^2} = -A\frac{k}{m}\sin\left(\sqrt{\frac{k}{m}}t\right) = -\frac{k}{m}x(t) \qquad (3.120)$$

with $v_{max} = A\omega$, and $a_{max} = A\omega^2$. For the momentum, $p(t) = mv$, we obtain:

$$p(t) = mv = m\frac{dx}{dt} = mA\sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}t\right) \qquad (3.121)$$

We can now solve for *E* exactly:

$$E = \frac{p^2}{2m} + \frac{kx^2}{2} \text{ or also } E = \frac{mv^2}{2} + \frac{kx^2}{2}$$

$$= \frac{\left[mA\sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}t\right)\right]^2}{2m} + \frac{k\left[A\sin\sqrt{\frac{k}{m}}t\right]^2}{2} \qquad (3.122)$$

Thus, the total mechanical energy of a simple oscillator is proportional to the square of the amplitude. As the amplitude (*A*) can take any value, this means that the energy (*E*) can also take any value—i.e., energy is continuous. Any energy value is allowed by simply changing the force constant *k*. At x = –A and x = A, $E = kA^2/2$, and at x = 0, $E = mv^2_{max}/2$ (see Figure 3.50).

In Figure 3.51 we show a typical nonquantized oscillator with a parabolic curve for the potential energy as a function of position $x - x_0$ of the mass. The potential sketched here is very important because it describes the potential for many systems, including vibrational and electronic states in molecules. It approximates the potential in many more systems for small departures from equilibrium. As a typical application, in Figure 3.52, we illustrate the potential energy for two hydrogen atoms approaching each other. Clearly over a considerable range of energies, a parabolic curve, typifying a simple harmonic oscillator, can represent the real situation. The horizontal lines cutting through the parabolic part of the curve in Figure 3.52 represent

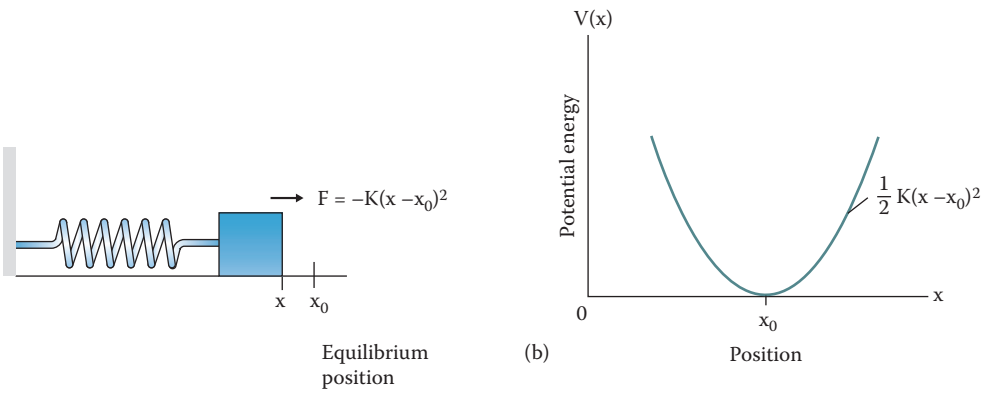


FIGURE 3.51 Typical nonquantized oscillator with a parabolic curve for the potential energy as a function of position *x* of the mass.

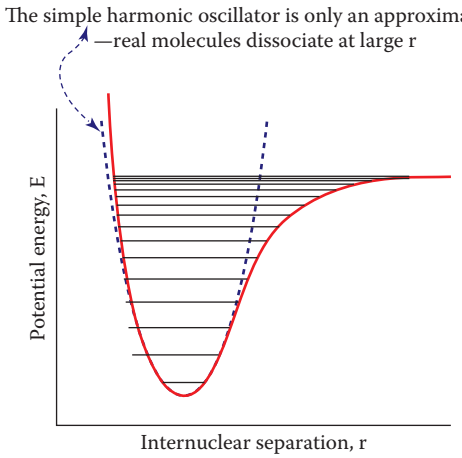


FIGURE 3.52 Application of simple oscillator approximation: two hydrogen atoms in a hydrogen molecule. For an explanation of the horizontal lines in this figure, see quantized oscillator below.

the quantization of the vibration states as introduced next.

The above examples demonstrate the essence of determinism in classical physics, i.e., given $V(x)$, one can solve for $x(t)$ or $v(x,t)$, or also at any time (t), the position $[x(t)]$ and velocity $[v(t)]$ can be determined exactly—i.e., the particle trajectory can be specified precisely. In other words, given the force, the motion can be found. In the late eighteenth century the mathematician Pierre Simon de Laplace (1749–1827) encapsulated classical determinism as follows: "…if at one time we knew the positions and motion of all the particles in the Universe, then we could calculate their behavior at any other time, in the past or the future." In classical physics, particles and trajectories are real entities, and it is assumed that the universe exists independently from the observer, that it is predictable, and that for every effect there is a cause so experiments are reproducible. Heisenberg's uncertainty principle destroyed all this. In quantum physics, measured and unmeasured particles are described differently. The measured particle has definite attributes such as position and momentum, but the unmeasured particle does not have one but all possible attribute values, as Nick Herbert describes it in his book *Quantum Reality* … somewhat like a broken television that displays all its channels at the same time.[3]

We shall see that these ideas of classical mechanics fail when we go to the atomic regime (where E and m are very small). Classical mechanics also fails when velocity is very large (as $v \rightarrow c$) because of relativistic effects.

Schrödinger's Equation

Plausibility of Schrödinger's Equation

From the evidence presented above, an atomic or subatomic particle cannot be described anymore as a simple Newtonian point, and *matter waves* must be taken into account. For matter waves, from $p = h/\lambda$ (Equation 3.95), the smaller the wavelength, the bigger the momentum, and as we calculated, electrons, for example, are very "wavy" particles. The book-keeping term for the energy of a system or the Hamiltonian, H, for a Newtonian particle with mass m is:

$$H = E = \frac{p^2}{2m} + V(x) \qquad (3.109)$$

An equivalent to Newton's equation is needed to calculate how forces (described by a potential energy) affect λ or $\mathbf{p}$ in the case of a "waving" particle. Once that differential wave equation is found, we may solve it for a wave, which has an amplitude for each value of position (x) and time (t).

Erwin Schrödinger (1887–1961), in 1926, encouraged by Debye, who remarked that there should be a wave equation to describe the de Broglie waves, proposed a wave equation that can be applied to any physical system in which it is possible to describe the energy mathematically. In one dimension he postulated:

$$\frac{\partial^2 \Psi(x,t)}{\partial x^2} + \frac{8\pi^2 m}{h^2}[E - V(x,t)]\Psi(x,t) = 0 \qquad (3.123)$$

with $\Psi(x,t)$ the wave function, a wave representing the spatial distribution of a "particle," and m the characteristic mass of the particle. The first term on the left is the rate of change of the wave function with distance x. The energy of the particle is E, and the potential energy function to describe the forces acting on the particle is represented by $V(x,t)$. Schrödinger's equation has the same central role in quantum mechanics that Newton's laws have in

mechanics and Maxwell's equations have in electromagnetism. Solutions to Newton's equations are of the form $v = f(x,t)$ (see Examples 3.1 and 3.2 above), whereas solutions to the wave equation (Equation 3.118) are called wave functions $\Psi(x,t)$. Schrödinger's equation is more difficult to solve than Newton's equation, but it is just as well-defined, and like Newton's equation, it describes the relation between kinetic energy, potential energy, and total energy. If one knows the forces involved, one can calculate the potential energy V and solve the equation to find Ψ. Solving Schrödinger's equation specifies $\Psi(x,t)$ completely, except for a constant; if $\Psi(x,t)$ is a solution, then $A\Psi(x,t)$ is a solution as well. Remember that Equation 3.123, like Newton's law, cannot be derived—it is a plausibility argument. Einstein called Ψ a *Gespensterfeld* or ghost field. Because it carries no energy, the wave function is also referred to as an empty wave. In France, the Ψ wave is sometimes called *densité de présence*, or *presence density*.

Here is how Schrödinger, using a rather sophisticated analogy with classical mechanics, came to derive Equation 3.123. He assumed a sinusoid wave of wavelength λ and frequency ν and hence a velocity $v = \lambda\nu$. The equation of such a traveling wave, as we saw in Chapter 2, is $\Psi(x,t) = A\cos(\mathbf{k}_x x - \omega t) + B\sin(\mathbf{k}_x x - \omega t)$, where $\mathbf{k}_x$ (the wave number) $= 2\pi/\lambda$, and the period is $T = 2\pi/\omega = 1/\nu$. Rewriting this expression in terms of the complex variable form, with $B = iA$ and using Euler's formula, one obtains as the simplest form of a wave:

$$\Psi(x,t) = Ae^{i(k_x x-\omega t)} \tag{3.124}$$

This is classical so far; indeed, we will see in the section "Solution of the Wave Equation for Free Particles", this chapter, and in Chapter 5 that Equation 3.124 is also a solution of the Maxwell equations.

Now apply Einstein's photon formula, i.e., $E = h\nu$, to the particle, with ν the frequency of the "waving" of the particle—de Broglie's idea! Because $E = h\nu = (h/2\pi)(2\pi\nu) = \hbar\omega$, we find $\omega = E/\hbar$, where we have introduced the definition of "h-bar," $\hbar = h/2\pi$. From de Broglie, $p = h/\lambda$ and $p = (2\pi/\lambda)(h/2\pi) = \mathbf{k}_x\hbar$ and $\mathbf{k}_x = p/\hbar$. Substitute the results for ω and $\mathbf{k}$ into Equation 3.124 and obtain:

$$\Psi(x,t) = Ae^{\frac{i(px-Et)}{\hbar}} \tag{3.125}$$

Schrödinger now assumed that the total energy E could be expressed in terms of the kinetic energy (KE) and the potential energy (PE) of the particle:

$$H = E = E_k + V(x) = \frac{p^2}{2m} + V(x,t) \tag{3.126}$$

The derivatives of Equation 3.125 are:

$$\frac{\partial\Psi(x,t)}{\partial x} = \frac{ip}{\hbar}\Psi(x,t)$$

so that

$$\frac{\partial^2\Psi(x,t)}{\partial^2 x} = \frac{-p^2}{\hbar^2}\Psi(x,t) \tag{3.127}$$

and

$$\frac{\partial\Psi(x,t)}{\partial t} = \frac{-iE}{\hbar}\Psi(x,t)$$

Note that Ψ cannot be canceled. Making use of the correspondence of the operators on $\psi(x,t)$ we can write: $\frac{\partial^2}{\partial^2 x} = \frac{-p^2}{h^2}$ and $\frac{\partial}{\partial t} = \frac{-iE}{h}$, and substituting these into the total energy equation (Equation 3.109) one derives:

$$E = i\hbar\frac{\partial}{\partial t} = -\left(\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\right) + V(x,t) \tag{3.128}$$

Applying these operators to $\Psi(x,t)$, we obtain:

$$E\Psi(x,t) = i\hbar\frac{\partial\Psi(x,t)}{\partial t} = -\left(\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2}\right) + V(x,t)\Psi(x,t) \tag{3.129}$$

which, the reader can easily demonstrate, is the famous Schrödinger's wave equation (Equation 3.123) for the 1D case. We may rewrite Equation 3.129 also in the form:

$$\left[-\frac{h^2}{8\pi^2 m}\frac{\partial^2}{\partial x^2} + V(x,t)\right]\Psi(x,t) = E\Psi(x,t) \tag{3.130}$$

This equation shows that H, the Hamiltonian operator, is given by:

$$H = -\frac{h^2}{8\pi^2 m}\frac{\partial^2}{\partial x^2} + V(x,t) \tag{3.131}$$

Thus, Equation 3.130 may simply be formulated in operator form as $H\Psi$ (operator Ψ acting on function

Ψ, an eigenfunction) = EΨ(function Ψ multiplied by a number E, an eigenvalue), where H is the 1D Hamiltonian operator and in which the energy E of the particles is called the eigenvalue, and Ψ the eigenfunction. Note again that Ψ cannot be canceled. The latter merely represents the wave associated with the particle. Expressed yet another way, kinetic and potential energies are transformed into the Hamiltonian, which acts on the wave function to generate the evolution of the wave function in time and space. Schrödinger's equation gives the quantized energies of the system and gives the form of the wave function so that other properties may be calculated.

The above does not represent a derivation of the wave equation; it is just a description of Schrödinger's thought process to make his postulate more plausible.

Wave Function Interpretation

Schrödinger did not have a clear idea of the meaning of a matter wave. However, in 1926 Max Born (1882–1970) presented a new, vivid interpretation of the particle wave function. The so-called "Copenhagen interpretation" of Schrödinger's equation is that the Ψ(x,t) function is not some physical representation of a physical substance as in classical physics (e.g., the amplitude of a water wave), but rather a "probability amplitude" of the particle, which, when squared, gives the probability of finding the particle at a given place at a given time: $|\Psi(x,t)|^2dx$ = probability the particle will be found between x and $x + dx$ at time t and the wave function itself has no physical meaning. Because Ψ(x,y,z,t) is complex and can be positive or negative, it cannot be the probability directly. The Born interpretation of Ψ places restrictions on the form of the wave function:

1. Ψ must be continuous (no breaks).
2. The gradient of Ψ($d\Psi/dx$) must be continuous (no kinks).
3. Ψ must have a single value at any point in space.
4. Ψ must be finite everywhere.
5. Ψ cannot be zero everywhere.

The Copenhagen interpretation also holds that an unmeasured particle is not real: its attributes are created or realized by the measuring act. Another way of saying this is that a wave function *collapses* up on measurement; before measurement a particle is described by a wave function described by Schrödinger's equation, but on measuring that particle's wave suddenly and discontinuously *collapses*. We will come back to this *mystic* interpretation of quantum reality when introducing *Schrödinger's cat* at the end of this chapter. Because the probability that the particle is somewhere must equal one, it holds that one can normalize this probability function as:

$$\int_{-\infty}^{+\infty} |\Psi(x,t)|^2 dx = 1 \qquad (3.132)$$

Note that the probability is a real number; although Ψ is complex, $|\Psi(x,t)|^2$ is real. In this case, Ψ is said to be a normalized wave function. Electrons do not fly around the nucleus like the Earth around the sun (Rutherford, Bohr), but depending on which energy level it is in, the electron can take one of a number of stationary probability cloud configurations (Schrödinger).

The boundary conditions imposed on Ψ mean that only certain wave functions and thus only certain energies of the system are allowed. Quantization of the wave function leads to quantization of the energy.

As we will learn below, solutions of Schrödinger's equation for an atom are spherical Bessel functions. In Figure 3.53 we show, as an example of the type

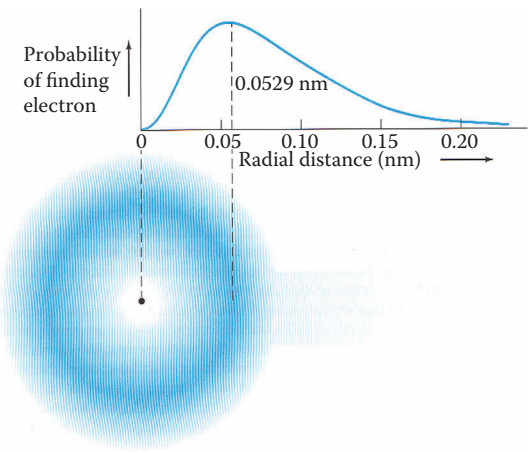


FIGURE 3.53 The probability of finding the ground state hydrogen electron (n = 1) as a function of the radial distance from the proton. The value of $|\Psi(x,t)|^2$ at some location is proportional to the probability of finding the particle at that location at that time.

of solutions obtained, the probability of finding an electron around the nucleus of a hydrogen atom. The potential that must be used in Schrödinger's equation for this case is $V(r) \propto 1/r$. Where we assume that the Coulomb force between the electron and the nucleus is the force responsible for binding the electron in the atom, this is the so-called central force or inverse square law $[F(r) \propto \frac{1}{r^2}]$.

Within Schrödinger's model the atom is regarded as a sort of vibrating balloon that extends to infinity and whose vibrations are in tune with Bohr's frequencies. The quantum numbers of Bohr and Sommerfeld are related to the number of nodes in this vibrating 3D system. This theory of matter waves also reproduces the Balmer series for the bright lines in the hydrogen atom. Schrödinger removed the mysterious discontinuous jumps between electron orbitals, replacing them with "beats" between the vibration frequencies of different quantum states. Schrödinger's elegant wave theory explains many things but lacks elsewhere, e.g., it cannot explain quantum processes such as the photoelectric effect.

The Time-Independent Schrödinger Equation (TISE) for Stationary States

The value of $|\Psi(x,t)|^2$ in Equation 3.132 at a particular point is in general a function of time. However, if the particle under consideration has a definite energy—think about an electron in a specific energy level in an atom—then the value of $|\Psi(x,t)|^2$ at each point becomes independent of time. It follows from quantum mechanics that for a particle in such a state of definite energy E, Ψ can be factored into a time-dependent component and a space-dependent component: $\Psi(x,t) = \psi(x)\psi(t)$:

$$\Psi(x,t) = \underset{\text{spatial}}{\psi(x)} \underset{\text{temporal}}{e^{-\frac{iEt}{\hbar}}} \qquad (3.133)$$

This is the time-dependent wave function for a stationary state where we use Ψ if the wave is a function of all coordinates and time, and ψ if it is a function of the space coordinates only. For simplicity we just deal with the x-coordinate here. The probability distribution function $|\Psi|^2$ for this state is the product of Ψ and its complex conjugate Ψ^*, or:

$$\begin{aligned} \left|\Psi(x,t)\right|^2 &= \Psi^*(x,t)\Psi(x,t) \\ &= \psi^*(x)\psi(x)e^{+\frac{iEt}{\hbar}}e^{-\frac{iEt}{\hbar}} \\ &= \psi^*(x)\psi(x)e^0 \\ &= |\Psi(x)|^2 \end{aligned} \qquad (3.134)$$

As $\psi(x)$ does not depend on time, we see from Equation 3.134 that the probability function does not depend on time either. As soon as one can define a state with a definite energy, one can define a stationary state.

We can now substitute the result from Equation 3.133 into the time-dependent Schrödinger's equation (Equation 3.129) to get:

$$i\hbar\psi(x)\frac{\partial\psi(t)}{\partial t} = -\left(\frac{\hbar^2}{2m}\psi(t)\frac{\partial^2\psi(x)}{\partial x^2}\right) + V(x,t)\psi(x)\psi(t) \qquad (3.135)$$

Dividing both sides by $\psi(x)\psi(t)$ results in:

$$i\hbar\frac{1}{\psi(t)}\frac{\partial\psi(t)}{\partial t} = -\left(\frac{\hbar^2}{2m}\frac{1}{\psi(x)}\frac{\partial^2\psi(x)}{\partial x^2}\right) + V(x,t) \qquad (3.136)$$

If the potential V is independent of t, then $V(x,t) = V(x)$, and the left side in Equation 3.136 depends on t only and the right side on x only. The only way these two sides can be equal to each other is if they are both equal to a constant, i.e., E. In the case $V(x,t)$ is independent of time, Equation 3.130 can thus be converted into a time-independent Schrödinger's equation (TISE). Hence we obtain the time-independent form of Schrödinger's equation as:

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x)\right]\psi(x) = E\psi(x) \qquad (3.137)$$

Solving this equation, say for an electron acted on by a fixed nucleus, we will see that this results in standing waves.

The more general Schrödinger equation features a time-dependent potential $V = V(x,t)$ and must be used, for example, when trying to find the wave function of an atom in an oscillating magnetic field or other time-dependent phenomena such as photon emission and absorption.

Applications of Schrödinger's Equation

Solution of the Wave Equation for Free Particles

The wave function of free particles, such as photons, phonons, plasmons, and "nearly free" particles, such as conduction electrons in metals, should all be solutions of Schrödinger's equation. In the case of freely traveling photons, the expression for planar light waves, $\Psi(x,t) = Ae^{i(k_x x-\omega t)}$ (1D case), derived from Maxwell's equations in Chapter 5, according to Bohr's correspondence principle, should also be a solution of Schrödinger's equation. To solve this first problem, we start from Equation 3.137, with V(x) = 0. This represents a free particle that experiences no force, has a definite energy E, and is moving in the x-direction with a momentum $\mathbf{p}_x$ (see Figure 3.54), say a photon:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = E\psi \text{ or } \frac{d^2\psi}{dx^2} = -\frac{2mE}{\hbar^2}\psi \qquad (3.138)$$

This is a second-order differential equation whose solutions are functions that, when differentiated twice, yield back the same functions multiplied by a constant. Such solutions include sines, cosines, and exponentials. Specifically, in this case, solving Equation 3.138 leads to the general solution, consisting of the superposition of two spatial wave functions of the form:

$$\psi(x) = Ae^{ik_x x} \qquad (3.139)$$

The two traveling waves, one traveling in the $+x$-direction and one traveling in the $-x$-direction, can be rewritten (using Euler) as:

$$\psi(x) = A_1[\cos(\mathbf{k}_x x) + i\sin(\mathbf{k}_x x)] + A_2[\cos(-\mathbf{k}_x x) + i\sin(-\mathbf{k}_x x)] \qquad (3.140)$$

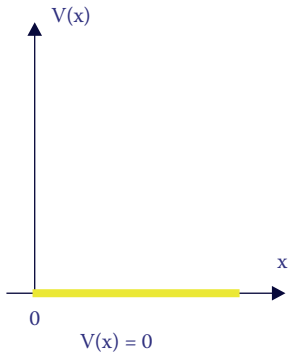


FIGURE 3.54 Photon, no boundary conditions, except for V(x) = 0.

where A_1 and A_2 are constants. We expected this result, of course, because in the preceding section we saw how Schrödinger built up his equation with the requirement that it would, at minimum, yield the same results that Maxwell's equations provide. This wave function for a free particle has a definite momentum $\mathbf{p}_x$ in the x-direction, or $\Delta\mathbf{p}_x = 0$. From the Heisenberg uncertainty principle (Equation 3.106), it then follows that Δx must be infinite; we have no idea where the particle is located in space. This makes sense as we are dealing with traveling waves. To confirm this further, we calculate the probability distribution function $|\Psi|^2$ for a free photon, which is the product of Ψ and its complex conjugate Ψ^*, or:

$$\begin{aligned}|\Psi(x,t)|^2 &= \Psi^*(x,t)\Psi(x,t)\\ &= (A^*e^{-ik_x x}e^{+i\omega t})(Ae^{ik_x x}e^{-i\omega t})\\ &= A^*Ae^0 = |A|^2\end{aligned} \qquad (3.141)$$

This result is independent of space and time, something we again expect for a sinusoidal wave function that extends all the way from $x = -\infty$ to $x = +\infty$ with an amplitude A. Normalization is not possible here as the wave extends to infinity; integration of Equation 3.139 over all space is infinite for any value of A.

By substituting $\psi(x) = Ae^{ik_x x}$ into Equation 3.138, we derive:

$$E = \mathbf{p}^2/2m \qquad (3.142)$$

and also:

$$\mathbf{k}_x = \frac{\sqrt{2mE}}{\hbar} \qquad (3.143)$$

With a positive $\mathbf{k}_x$, the wave function represents a free particle moving in the positive x-direction. If $\mathbf{k}_x$ is negative, the motion is in the negative x-direction. For a free particle, there is no restriction on the value of $\mathbf{k}_x$, and the associated, unquantized energy from Equation 3.143 is given as:

$$E(\mathbf{k}) = \frac{\hbar^2\mathbf{k}_x^2}{2m} \qquad (3.144)$$

Any positive value of energy is allowed for such a free wave/particle; there is nothing that restricts the values of E. Equation 3.144 is the so-called dispersion relation for free particles. When we plot the

energy E versus the wave vector $\mathbf{k}_x$ for free particles, we obtain the parabola shown in Figure 3.55. It is appreciated from this plot that mass, m, is inversely proportional to curvature.

Free Electrons in an Infinite Piece of Metal

For free electrons in an infinitely large 3D piece of metal, the allowed electron states are solutions of an expanded version of Schrödinger's equation in Equation 3.123, or:

$$\nabla^2\psi(\mathbf{r}) + \frac{8\pi^2 m}{h^2}[E - V(\mathbf{r})]\psi(\mathbf{r}) = 0 \quad (3.145)$$

For electrons swarming around freely in this infinite metal, the potential energy $V(r)$ is zero inside the conductor, and the solutions inside the metal are plane waves moving in the direction of $\mathbf{r}$:

$$\Psi_k(\mathbf{r}) = A\exp i\mathbf{k}\cdot\mathbf{r} \quad (3.146)$$

where $\mathbf{r}$ is any vector in real space, and $\mathbf{k}$ is any wave vector. As with a freely moving particle, normalization is impossible as the wave extends to infinity.

Plotting the energy E versus the wave number $\mathbf{k}_x$ for a free electron gas in one direction leads to the same parabolic dispersion relation shown in Figure 3.55. The density of states (DOS) function $G(E)$ is the number of possible energy states per unit volume (also degeneracy), and we will derive it further below for free electrons in a bulk piece of solid. We find that, just as in the case of the density of states of an ideal gas (Equation 3.19), it increases smoothly with the square root of the energy. Importantly, we will also find that when reducing the dimensionality of the solid, say to a plane, a line, or a dot, the density of state function changes dramatically and acquires more density of states at specific energy values; i.e., $G(E)$ is not a smooth function of E anymore.

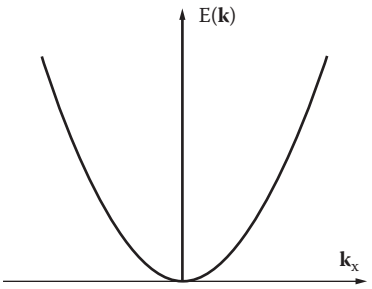


FIGURE 3.55 Plot of energy versus wave number $\mathbf{k}_x$ from Equation 3.144 for free particles. Dispersion relation $E(\mathbf{k}_x)$.

From classical theory we could not appreciate the occurrence of long electronic mean free paths; indeed, Drude used the interatomic distance a for the mean free path λ. But from experiments with very pure materials and at low temperatures it is clear the mean free path may be much longer; actually it may be as long as 10^8 or 10^9 interatomic spacings or more than 1 cm. The quantum physics answer is that the conduction electrons are not deflected by ion cores arranged in a periodic lattice because matter waves propagate freely through a periodic structure just as predicted by Equation 3.146.

Confining electrons by limiting their propagation in certain directions in a crystal introduces a varying $V(r)$ in Schrödinger's equation, and this may lead to an electronic bandgap as we will introduce below.

Particles in Infinitely Deep Potential Wells of Finite Size

The Born and von Karman's Periodic Boundary Condition According to Pauli's exclusion principle, in an atom, no two electrons can have all four quantum numbers the same. We ask ourselves now if there are similar restrictions for electrons in a larger structure. Instead of the infinitely large piece of metal, considered above, we limit the size of the metal chunk to say a 1-cm^3 piece of metal, and we find that restrictions for electron energies do indeed materialize. Discrete energy levels inevitably arise whenever a small particle such as a photon or electron is confined to a region in space. Sommerfeld, in 1928, was the first to show this. He adopted Drude's free electron or Fermi gas (FEG) and added the restriction that the electrons must behave in accordance with the rules of quantum mechanics. In his Fermi gas, electrons are free, except for their confinement within a cubic piece of crystalline conductor with a volume of $V = L^3$, and they follow Fermi-Dirac statistics instead of Maxwell-Boltzmann rules. The choice of a cube shape is a matter of mathematical convenience; a periodic boundary condition ensures that the free electron form of the wave function is NOT modified by the shape of the conductor or its boundary. This can be interpreted as follows: an electron coming to the surface is not reflected back in but reenters the metal piece from the opposite surface. This excludes the surfaces from playing any

role in transport phenomena. The value L is set by the Born and von Karman's periodic boundary condition, i.e., that the wave functions must obey the following rule:

$$\psi(x + L, y + L, z + L) = \psi(x,y,z) \qquad (3.147)$$

For the derivation of the possible energies, we assume N electrons (one for each metal ion) in a cube of solid conductor with sides of length L.

Outside the 3D cube of solid the potential $V \rightarrow \infty$, and the wave function ψ is zero anywhere outside the solid with $x,y,z \leq 0$ and $x,y,z \geq L$. The situation applies, for example, to totally free electrons in a metal where the ion cores do not influence their movement. Sommerfeld assumed that $V(x)$ outside the conductor equaled the work function Φ. The work function is the amount of energy required to remove an electron from the surface of a metal, i.e., the height of the wall electrons would have to scale to escape the solid, but $V = \infty$ is a good enough approximation for electrons in low-energy levels. Before we apply the Born and von Karman's periodic boundary condition to a finite-size 3D box that is infinitely deep ($V = \infty$), we consider finite-size infinitely deep 1D and 2D wells.

Infinitely Deep, Finite-Sized 1D Potential Wells—Quantum Wells We apply Equation 3.137 now to a finite-sized 1D box with infinitely high potential walls (Figure 3.56). In a 1D box, $V(x)$ in Equation 3.137—the time-independent Schrödinger equation (TISE)—is 0 everywhere inside the conductor (region II) and is infinite outside the conductor (regions I and III).

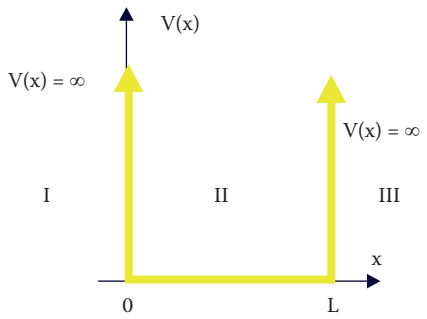


FIGURE 3.56 Electron in a box. L is not the real physical boundary of the conductor, as the surfaces of the conductor are not determining the physical properties. The value of L is set by the Born and von Karman periodic boundary condition (see Equation 3.147).

Outside the well $V \rightarrow \infty$ and $\psi = 0$ for $x \leq 0$ and $x \geq L$. For the Schrödinger equation inside the well ($0 < x < L$), we write:

$$-\frac{\hbar^2}{2m_e}\frac{d^2\psi(x)}{dx^2} = E\psi(x) \qquad (3.148)$$

or:

$$\frac{d^2\psi}{dx^2} = -k_x^2\,\psi \qquad (3.149)$$

with:

$$k_x = \sqrt{\frac{2m_eE}{\hbar^2}} \qquad (3.150)$$

We are interested in the quantum mechanically allowed energy levels of the electrons in this 1D box; because $V(x) = 0$, these will be kinetic energy levels (from Equation 3.109, $E = p^2/2m_e$). From Equation 3.148 the kinetic energy is proportional to the curvature of ψ (the curvature of a function is its second derivative).

In Sommerfeld's mathematical model, for $x = 0$ and at $x = L$, the wave function must be zero. The general stationary state solution for Equation 3.149 consists of the superposition of two spatial wave functions $\psi(x)=Ae^{ik_xx}$, as shown in Equation 3.140, which we rewrite here as:

$$\psi(x) = (A_1 + A_2)\cos k_x x + i(A_1 - A_2)\sin k_x x \qquad (3.151)$$

with one traveling in the $+x$-direction and one traveling in the $-x$-direction. The boundary conditions are set as follows: $x = 0$, $\psi(x) = A_1 + A_2 = 0$ or $A_2 = -A_1$, and we rewrite Equation 3.151 as:

$$\psi(x) = 2iA_1 \sin k_x x = \psi_0 \sin k_x x \qquad (3.152)$$

From the second boundary condition that $\Psi_0 \sin(k_xL)$ must be 0, it follows that either $\psi_0 = 0$ or $\sin(k_xL) = 0$. If $\psi_0 = 0$, the wave function is zero everywhere, or there is no probability of finding the particle in the box anywhere; therefore, this solution must be disregarded. Hence $k_xL = n\pi$, where $n = 1, 2, 3,\ldots$ or:

$$\psi(x) = \psi_0 \sin\left(\frac{n\pi x}{L}\right) \qquad (3.153)$$

This is a standing wave that acts like a string clamped down at both ends. With n = 1 one has the basic harmonic, and with n = 2, 3,... higher harmonics result. Note that the constant, ψ_0, cannot be determined from solving Schrödinger's equation because we are dealing with a linear equation. The constant must be recovered from the normalization condition, which is required to give a probability interpretation to the wave function. The constant ψ_0 is calculated as:

$$\int_{-\infty}^{+\infty} |\Psi(x,t)|^2 \, dx = 1 = |\psi_0|^2 \int_0^L \sin^2\left(\frac{n\pi x}{L}\right) dx = |\psi_0|^2 \frac{L}{2} \tag{3.154}$$

or we find:

$$\Psi_n(x,t) = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \tag{3.155}$$

with n = 1,2,3,…. These are the eigenfunctions of an electron in a box. The discrete energies—energy eigenvalues—that the electron can adopt are:

$$E_n(\mathbf{k}) = \frac{\hbar^2}{2m_e}\left(\frac{n\pi}{L}\right)^2 \tag{3.156}$$

The values of $E_n(\mathbf{k})$ for different quantum numbers *n* represent the various allowed energy levels in a 1D box, and the gap between two successive levels describes the effect of quantization (discreteness). At the lowest energy (n = 1), the *ground state*, the energy remains finite despite the fact that V = 0 inside the region. According to quantum mechanics, an electron cannot be inside the box and have zero energy. This is called the zero-point energy, an important consequence of Heisenberg's uncertainty principle (we elaborate on this point on p. 114 "Heisenberg's Uncertainty Priniciple"). If V(x) ≠ 0, everywhere in the box, all energies are shifted by V ($E_n = \frac{n^2h^2}{8m_eL^2} + V$). For the same value of quantum number *n*, the energy is inversely proportional to the mass of the particle and to the square of the length of the box. For a heavier particle and a larger box, the energy levels become more closely spaced. Only when m_eL^2 is of the same order as h^2 do quantized energy levels become important in experimental measurements (with L = 1 nm, $E_1 = \frac{h^2}{8m_eL^2} = 0.36$ eV). With a 1-cm^3 piece of metal (instead of 1 nm^3), the energy levels become so closely spaced that they seem to be continuous ($E_1 = \frac{h^2}{8m_eL^2} = 3.6 \times 10^{-15}$eV); in other words, the quantum mechanics formula gives the classical result for dimension such that $m_eL^2 >> h^2$. This is another nice illustration of Bohr's correspondence principle: for large dimensions, Schrödinger's equation yields the classical results back. Because the kinetic energy is proportional to the curvature of the wave function, a higher kinetic energy E_n, caused by a higher value of *n* and/or a smaller value for *L*, corresponds to a more curved wave function (shorter wavelength). On a molecular scale, because L^2 appears in the denominator of Equation 3.156, increasing *L*, the size of the box, stabilizes the particle. Chemists are familiar with this effect in the case of electrons where delocalization is a stabilizing factor. As a consequence, allyl and benzyl carbonium ions, radicals, and carbanions are relatively stable.

The wave number $\mathbf{k}_x$ given as:

$$\mathbf{k}_x = \frac{2\pi p}{h} = \frac{p}{\hbar} = \frac{\sqrt{2m_eE_n}}{\hbar} \tag{3.150}$$

is now also quantized because E_n is quantized so that $\mathbf{k}_x = n\pi/L_x$ where *n* can take on only values 1, 2, 3….

Quantization occurs because of boundary conditions and the requirement for ψ to be physically reasonable (Born interpretation). The quantum number *n* labels each allowed state (ψ_n) of the system and determines its energy (E_n). Knowing *n*, we can calculate ψ_n and E_n. The wave number is related to the momentum $\mathbf{p}_x$ of the electron, viewed as a particle, by $\mathbf{k}_x = 2\pi\mathbf{p}_x/h$, and to the wavelength (λ) of the electron, viewed as a wave, by $\mathbf{k}_x = 2\pi/\lambda$. The dispersion function $E(\mathbf{k}_x)$ for an electron in an infinite deep well of finite size is shown in Figure 3.57. The dispersion or energy spectrum $E(\mathbf{k}_x)$ is now discrete rather than continuous; the allowed wave vectors are uniformly spaced in **k**-space with a separation of π/L; and the sample size *L* determines the spacing of allowed wave vectors and single-particle energies, with a smaller box giving

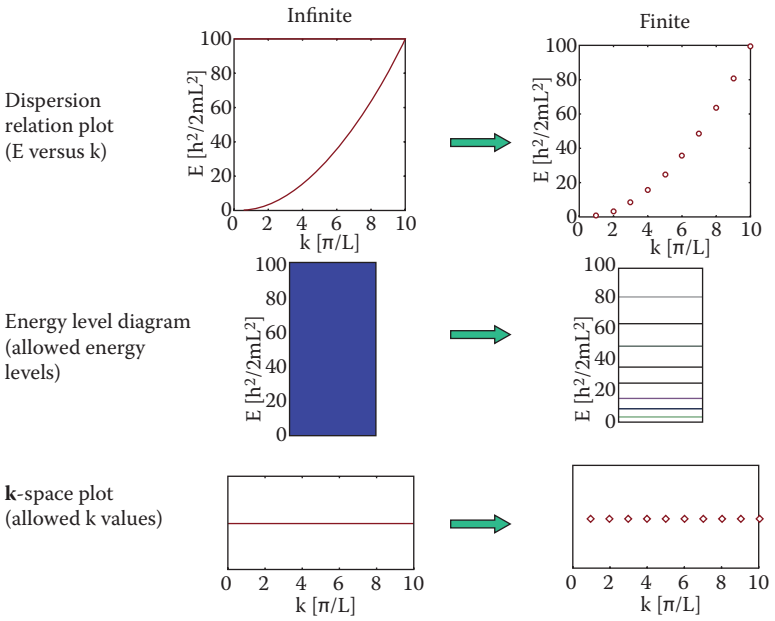


FIGURE 3.57 Dispersion $E(\mathbf{k}_x)$ function for an electron in an infinite box compared with the dispersion of the electron in a finite-size box. Finite size drastically alters allowed energy levels.

the larger spacings. For the free electron, no such values exist (see Figure 3.55).

The ground state (n = 1) of the wave function ψ_1 (Equation 3.150) has no "zero crossings." The first excited state (n = 2) has one zero crossing, and so on. Successive functions possess one more half-wave (they have a shorter wavelength). Nodes in the wave function are points at which $\psi_n = 0$ (excluding the ends, which are constrained to be zero), and the number of nodes = n − 1. The important point to notice is that the imposition of the boundary conditions has restricted the energy to discrete values. The allowed wave functions, a family of standing waves, and energies of an electron in a 1D box are summarized in Figure 3.58.

As apparent from Figure 3.58, the probability distribution of the particle in a box is not uniform; however; for a very large n, the probability is almost uniform throughout the box, as dictated again by classical physics. The probability of finding the particle at some point varies with the energy of the particle. A particle with energy E_1 is most likely to be found in the middle of the box. A particle with energy E_2 will never be found at that spot. As the energy of the particle increases, so does the number of nodes in the eigenfunction. Increasing the number of nodes (decreasing λ) corresponds to increasing kinetic energy.

Devices that come with a length L in one direction comparable with the size of the de Broglie wavelength of an electron are known as quantum wells (1D confinement). With at least one dimension between 1 and 100 nm, the excess electrons in this confined direction have no room to move in a Newtonian fashion. Instead their positions and velocities take on a probabilistic nature as their wave nature now dominates. Charge carriers are still free to move in the other two unconstrained directions though and

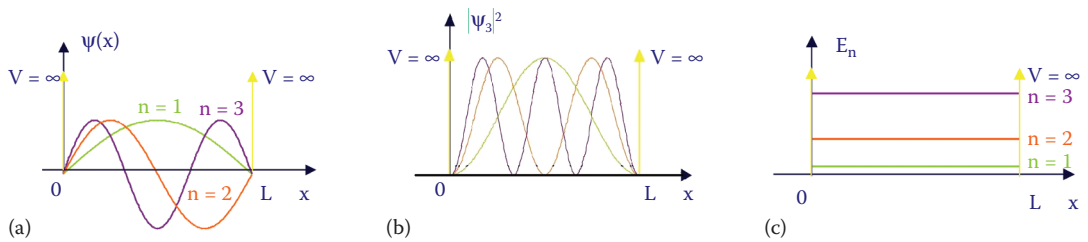


FIGURE 3.58 Particle in a box: overview. (a) Standing waves in the box. (b) The probability distribution $|\psi_3|^2$. (c) The energy is proportional to n^2.

form a 2D carrier gas in a plane perpendicular to the confinement direction (the *x*-*y* plane if the confinement is in the *z*-direction). Quantum wells are relatively easy and cheap to fabricate. By using two different types of semiconductor materials, one of these having a wider bandgap, quantum wells are formed on sandwiching a thin layer of a narrow bandgap material between wider bandgap semiconductor layers. This is realized, for example, by sandwiching a thin epitaxial layer of GaAs between two layers of $Al_xGa_{1-x}As$, in a so-called heterostructure as depicted in Figure 3.59. The narrower bandgap GaAs is enclosed by $Al_xGa_{1-x}As$, a material with a considerably larger bandgap to establish a potential barrier for electrons at the surface of the confined material. The motion of electrons and holes is thereby restricted in the direction perpendicular to the thin layer of GaAs (*z*-direction). In this structure, electrons are still free to move unrestricted in the *x*-*y* plane.

These quantum wells (QWs) were developed in the early 1970s and constituted the first lower dimensional heterostructures. The foremost advantage of such a design involves their improved optical properties. Reduced dimensionality leads to marked improved optical performances because of the change in the density of state or DOS function (see p. 161 "Fermi Surfaces, Brillouin Zones, Density of State Functions, and Conductivity as a Function of Quantum Confinement" for more details) for such a device compared with its 3D counterpart. In contrast to a bulk semiconductor, in a QW there are no allowed electron states at the very lowest energies (an electron in a box with energy = 0 does not exist; see p. 125 "Infinitely Deep, Finite-Sized 1D Potential Wells—Quantum Wells"), but there are many more available states (higher DOS) in the lowest conduction state in a QW so that many more electrons can be accommodated. Similarly, the top of the valence band has plenty more states available for holes. This means that it is possible for many more holes and electrons to combine and produce photons with identical energy for enhanced probability of stimulated emission (lasing) (see Chapter 5). Optical properties can be tuned by changing the structural parameters of a QW, principally its thickness and composition; this is known as bandgap engineering.

Today QWs form the basis of most optoelectronic devices. QWs are used, for example, in high electron mobility transistors (HEMTs) and solar cells with high efficiency. HEMTs are in use in all types of high frequency electronics such as cell phone and satellite television. The reason for the higher speed of a HEMT over a classical transistor is that the mean free path of the charge carriers in the 2D GaAs layer of a HEMT can be made larger than the gate length of the transistor, resulting in ballistic transport, i.e., charge transport without any intervening collisions with other charge carriers. Another mature application of QWs involves solid-state lasers. With the right voltages applied over a single QW, very large numbers of holes and electrons can be brought together in a tiny physical space and narrow energy range, leading to powerful surface emitting lasers as used in $5 laser pointers, compact disc players, and laser printers.

Almost 30 years after quantum wells were first developed, the 2000 Nobel Prize in Physics was awarded to Zhores Alferov and Herbert Kroemer (Figure 3.60) for their contributions in the field of semiconductor heterostructures and their high-speed and optoelectronics applications.

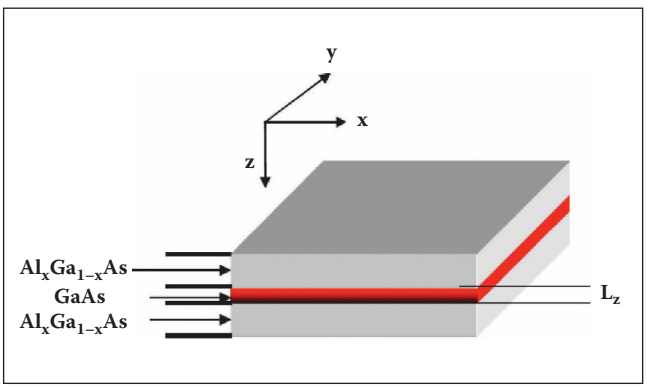


FIGURE 3.59 Quantum well with *x*,*y* dimensions infinite and L_z finite.[4]

FIGURE 3.60 Zhores Alferov (a) and Herbert Kroemer (b) were awarded the Nobel Prize in Physics in 2000.

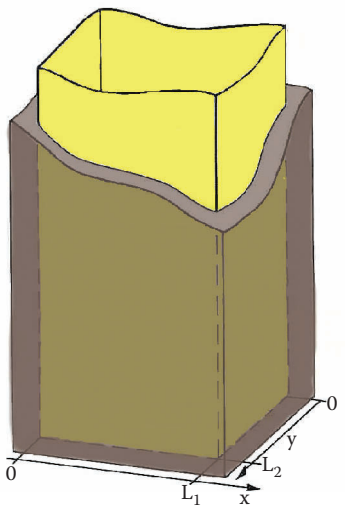


FIGURE 3.61 A 2D square well. A quantum wire.

Led by the insights garnered on QWs, scientists investigated the possibility of reducing the dimensionality of heterostructures even further to create 1D (quantum wire) and 0D (quantum dot) structures (see following sections).

Infinitely Deep, Finite-Size 2D Potential Wells—Quantum Wires For a particle in a finite-sized, 2D infinitely deep potential well, we define a wave function similar to the 1D potential well, but now we obtain $\psi(x,y)$ solutions that are defined by two quantum numbers, one associated with each confined dimension. The pertinent wave functions and energies for a 2D infinitely deep potential well, as shown in Figure 3.61, are:

$$\Psi_{n_1n_2}(x,y) = \left(\frac{4}{L_1L_2}\right)^{\frac{1}{2}} \sin\left(\frac{n_1\pi x}{L_1}\right)\sin\left(\frac{n_2\pi y}{L_2}\right) \quad (3.157)$$

and:

$$E_{n_1n_2}(\mathbf{k}) = E^x_{n_1} + E^y_{n_2} = \frac{h^2}{8m}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right) \quad (3.158)$$

with $n_1 = 1, 2, 3$ and $n_2 = 1, 2, 3$ and where L_1 is along the x-axis and L_2 along the y-axis, making the z-direction the unconstrained axis. The energy is quantized now along the x- and y-axes as a consequence of the boundary conditions, and most of the features of the 1D well are reproduced. Some of the low-energy wave functions together with their contour maps are illustrated in Figure 3.62.

One new feature in 2D potential wells, not found in 1D potential wells, is degeneracy, i.e., the occurrence of several quantum states at the same energy level. This is best understood through an inspection of Equations 3.157 and 3.158: for a square well instead of a rectangular one, $L_1 = L_2 = L$, and

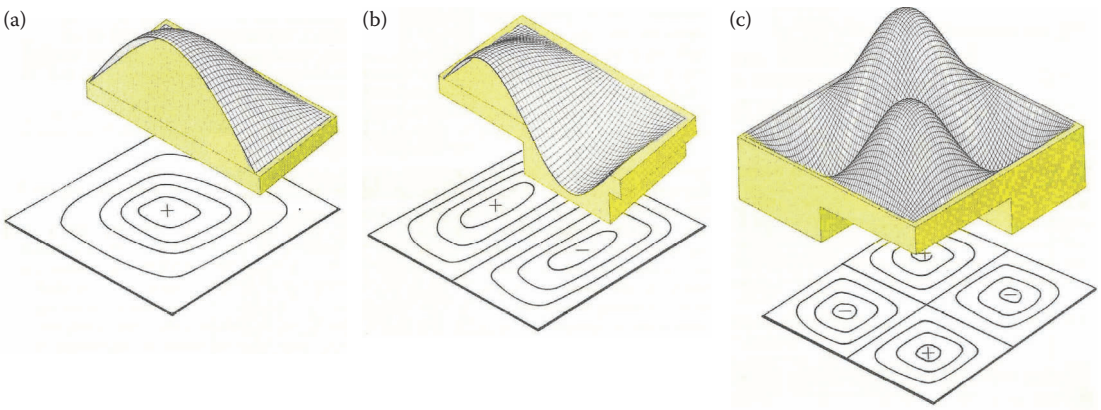


FIGURE 3.62 (a) ψ_{11}, (b) ψ_{21}, and (c) ψ_{22} for a 2D square potential well. A quantum wire.

Equation 3.158 yields now the same energy of $E_{n_1n_2}(\mathbf{k}) = E^x_{n_1} + E^y_{n_2} = \frac{h^2}{8mL^2}(n_1^2 + n_2^2)$. A state with $n_1 = b$ and $n_2 = a$ has the same energy as a state with $n_2 = a$ and $n_2 = b$ (even if $a \neq b$); these different states that correspond to a same energy are called degenerate. That these are indeed different quantum states becomes obvious when substituting the quantum numbers for such states in Equation 3.157. In the case of two degenerate states with $a = 1$ and $b = 2$, both corresponding to an energy of $\frac{5h^2}{8mL^2}$, we obtain two different wave functions:

$$\Psi_{1,2}(x,y) = \left(\frac{2}{L}\right)\sin\left(\frac{\pi x}{L}\right)\sin\left(\frac{2\pi y}{L}\right) \quad \text{and}$$

$$\Psi_{2,1}(x,y) = \left(\frac{2}{L}\right)\sin\left(\frac{2\pi x}{L}\right)\sin\left(\frac{\pi y}{L}\right).$$

Degeneracy is obviously connected to the degree of symmetry of a system; in the case we make an asymmetric box with $L_1 \neq L_2$ (rectangle instead of a square), the degeneracy disappears.

When *L* is very small along two directions (2D confinement)—of the order of the de Broglie wavelength of an electron—one obtains a quantum wire where electrons can only move freely in one direction, i.e., along the length of the quantum wire as illustrated in Figure 3.63.

Examples of 2D quantum confinement comprise nanowires and carbon nanotubes. Quantum wires represent the smallest dimension feasible for efficient transport of electrons and are thus aimed at as the ultimate interconnects in nanoelectronics and nanooptoelectronics. We mentioned above that quantum well lasers are superior over traditional-bulk solid-state lasers. Structures with yet lower dimensionality, such as nanowires and quantum dots (see next section), are even better, coming with a lower threshold current and switching on and off faster than quantum well lasers (~40 GHz and higher). Quantum wires also have been made into transistors (bipolar and FET), inverters, LEDs, and memory structures.

Compared with the fabrication of quantum wells, the realization of nanoscale quantum wires requires more difficult and precise growth control in the lateral dimension, and, as a result, quantum wire applications are only in the development stage. Quantum wire fabrication techniques include nanoscale lithography, self-organization, selective growth, and chemical and electrochemical synthesis (see Volume III, Chapter 3). In the top-down approach, taking advantage of well-developed quantum well fabrication technologies, using molecular beam epitaxy (MBE) and metalorganic vapor deposition (MOCVD), the most straightforward method to realize 1D nanostructures is etching (and regrowth) through wire-defining masks placed above a quantum well. This way, GaAs quantum wires are fabricated starting from the same thin layer of GaAs sandwiched between two layers of Al_xGa_{1-x} As we encountered in the manufacture of quantum wells. In the bottom-up approach, quantum wires are formed via direct growth in the form of semiconductor or metal nanowires and carbon nanotubes. Sumio Iijima (Figure 3.64) discovered multiwall carbon nanotubes (MWNTs) in 1991, after experimenting with

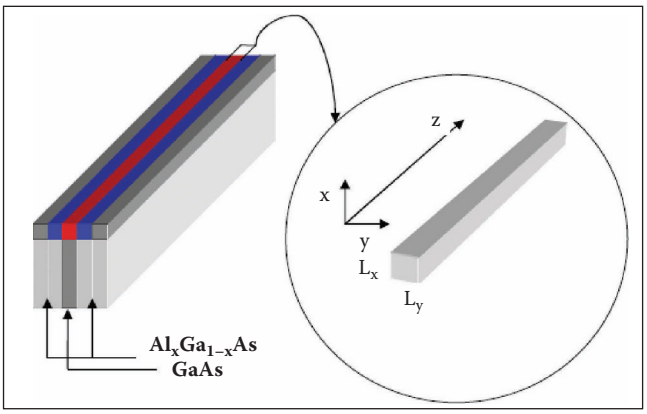


FIGURE 3.63 Quantum wire with dimensions *z* infinite and L_x, L_y finite.[4]

FIGURE 3.64 NEC's Sumio Iijima discovered multiwall carbon nanotubes (MWNTs) in 1991.

an arc-discharge technique similar to the one used by Richard Smalley and his team at Rice University in the discovery of buckminsterfullerene (C60). During arc discharge between two closely spaced graphite rods, carbon vaporizes and after condensing yields a "sooty" mass. When Iijima looked at the soot with an SEM, he noticed hollow nanotubes of carbon. One may speculate that this finding, like that of C60 "buckyballs," could have been achieved earlier if better microscopy techniques had been available to see the nano-sized products hidden in the soot.

During the past decade, there has been major progress reported in the chemical synthesis (i.e., bottom-up manufacture) of all types of nanoscale semiconductor wires. As originally proposed by R.S. Wagner and W.C. Ellis, from Bell Labs, for the Au-catalyzed Si whisker growth, a vapor-liquid-solid mechanism is still mostly used.[5] The field got a shot in the arm (a rebirth, so to speak) with efforts by Charles Lieber (Harvard) (Figure 3.65), Peidong Yang (http://www.cchem.berkeley.edu/pdygrp/main.html), James Heath (http://www.its.caltech.edu/~heathgrp), and Hongkun Park (http://www.people.fas.harvard.edu/~hpark). Lieber's group at Harvard (http://cmliris.harvard.edu) reported arranging indium phosphide semiconducting nanowires into a simple configuration that resembled the lines in a tick-tack-toe board.

The team used electron beam lithography to place electrical contacts at the ends of the nanowires to show that the array was electronically active. The tiny arrangement is not yet a circuit, but it is a first step, showing that separate nanowires can be contacted to one another.

Infinitely Deep, Finite-Sized 3D Potential Wells—Quantum Dots The solution of Schrödinger's equation for electrons in a cube of metal with side L is given as:

$$\Psi_k(\mathbf{r}) = V^{-\frac{1}{2}}\exp(i\mathbf{k}\cdot\mathbf{r}) \qquad (3.159)$$

—a 3D generalization of Equation 3.157 (with $V = L^3$). The wave vector $\mathbf{k}$, if you recall, points in the direction of wave propagation and has a magnitude given by $2\pi/\lambda$ for a plane wave. Generally, if a wave propagates along a displacement vector $\mathbf{r}$, the

FIGURE 3.65 Harvard's Charles Lieber reinvigorated the nanowire field.

amount of phase accumulated by the wave is given by $\mathbf{k}\cdot\mathbf{r}$. From the generalization of Equation 3.158, electrons in a cubic box of side L ($L_1 = L_2 = L_3$) have allowed energy levels specified by three quantum numbers n_1, n_2, and n_3, or:

$$E_{n_1, n_2, n_3} = \frac{h^2}{8m}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2} + \frac{n_3^2}{L_3^2}\right)$$

$$= \frac{h^2}{8mL^2}(n_1^2 + n_2^2 + n_3^2) \qquad (3.160)$$

and they come with the following allowed wave vectors:

$$\mathbf{k}_x = \frac{2\pi n_x}{L}, \mathbf{k}_y = \frac{2\pi n_y}{L}, \mathbf{k}_z = \frac{2\pi n_z}{L},$$

$$n_x, n_y, n_z = 1, 2, 3, \ldots \text{(integers)} \qquad (3.161)$$

The allowed $\mathbf{k}$-states are uniformly spaced along the axes with one state per length $2\pi/L$. This is shown for a 2D square array and a 3D cubic array in Figure 3.66. One consequence of confining a quantum particle in a cubic 3D is again "degeneracy." As we

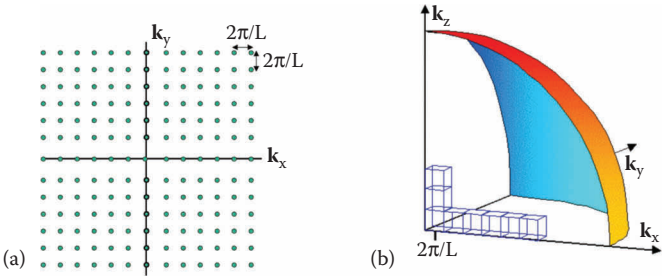


FIGURE 3.66 A 2D array of allowed $\mathbf{k}$-state where we are looking in "$\mathbf{k}$-space" or "reciprocal space" (so-called because $\mathbf{k}$ has units of $1/L$) (a) and a cubic array of allowed states in 3D $\mathbf{k}$-space; in three dimensions, we have states described by $\mathbf{k} = (\mathbf{k}_x; \mathbf{k}_y; \mathbf{k}_z)$ (b). The area per point is $(2\pi/L)^2$, and the volume per point is $(2\pi/L)^3$.

saw for the 2D case, degeneracy reflects an underlying symmetry in the $V(x,y,z)$ potential and can be removed if certain symmetry is broken (e.g., by making $L_1 \neq L_2 \neq L_3$). Each dot in Figure 3.66 represents an allowed **k**-state. An important feature here is that the larger the box, the closer the spacing in energy of single particle states, and vice versa. The **k**-states are discrete, but for a normal-sized conductor there are $\sim 10^{26}$ states, so we can treat them as a continuum.

From Figure 3.66a, each value of **k** occupies an area $A = (2\pi/L)^2$, and from Figure 3.66b each point occupies a volume $V = (2\pi/L)^3$. The number of states per unit volume of **k**-space is $1/V$ (or $L/2\pi)^3$.

When L becomes very small, of the order of the de Broglie wavelength of an electron in all three directions, the electrons lose all capacity to move (3D confinement). These 0D structures, with 3D quantum confinement, are called quantum dots or QDs, as illustrated in Figure 3.67. Quantum dots, also known as nanocrystal semiconductors, ranging from 2 to 10 nm (10–50 atoms) in diameter, are typically composed of materials from periodic groups II–VI, III–V, or IV–VI (e.g., CdS, CdSe, PbS, PbSe, PbTd, CuCl…). The trapped electrons in these dots behave as de Broglie standing waves. The confinement of the waves leads to a blue energy shift, and by varying the particle size one can produce any color in the visible spectrum, say from deep (almost infra-) reds to screaming (almost ultra-) violets as illustrated in Figure 1.31. A quantum dot (QD) is an atom-like state of matter sometimes referred to as an "artificial atom." What is so interesting about a QD is that electrons trapped in them arrange themselves as if they were part of an atom, although there is no nucleus for the electrons to surround here. The type of atom the dot emulates depends on the number of atoms in the well and the geometry of the potential well $V(x)$ that surrounds them.

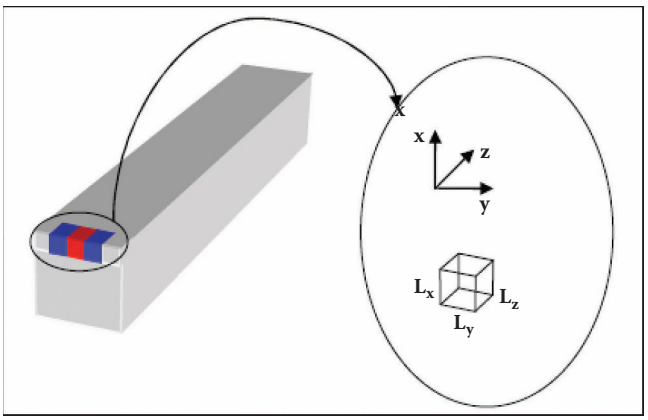


FIGURE 3.67 Quantum dot with dimensions L_x, L_y, and L_z that are finite.[4]

An important consequence of decreasing the dimensionality even further than in quantum wells and wires is that the density of state function for quantum dots features an even sharper and yet more discrete density of states. As a consequence, quantum dot lasers exhibit a yet lower threshold current than lasers based on quantum wires and quantum wells, and because of the more widely separated discrete quantum states, they are also less temperature sensitive. However, because the active lasing material volume is very small in quantum wires and dots, a large array of them has to be made to reach a high enough overall intensity. Making quantum wires and dot arrays with a very narrow size distribution to reduce inhomogeneous broadening remains a real manufacturing challenge, and as a result only quantum well lasers are commercially mature.

However, quantum dots already form an important alternative to organic dye molecules. Unlike fluorescent dyes, which tend to decompose and lose their ability to fluoresce, quantum dots maintain their integrity, withstanding many more cycles of excitation and light emission (they do not bleach as easily!). Combining a number of quantum dots in a bead conjugated to a biomolecule is used as a spectroscopic signature—like a bar code on a commercial product—for tagging those biomolecules (see Chapter 7, "Fluorophores, Quantum Dots, and Fluorescent Proteins").

In the early 1980s, Dr. Ekimov discovered quantum dots with his colleague, Dr. Efros, while working at the Ioffe Institute in St. Petersburg (then Leningrad), Russia.[6,7] This team's discovery of quantum dots occurred at nearly the same time as Dr. Louis E. Brus (Figure 3.68), a physical chemist then working at AT&T Bell Labs (he is now at Columbia), found out how to grow CdSe nanocrystals in a controlled manner.[8,9] Experimenting with CdSe nanocrystal semiconductor material, Dr. Brus and his collaborators observed solutions of astonishingly different colors made from this same substance. Their observation led to the recognition that there is a very clear transition in material behavior when

FIGURE 3.68 Columbia's Louis Brus.

a chunk of it becomes smaller than a fundamental scale, intrinsic to the substance.

Even though it was predicted in 1982 that QDs could be used as the active region of lasers, providing reduced threshold current and lower temperature dependence, it took nearly a decade to develop reliable growth techniques to produce QDs of a quality suitable for commercial applications. Major contributions to making QDs a reality were made by two Bell Labs scientists, Dr. Moungi Bawendi and Dr. Paul Alivisatos. They have since moved to MIT and University of California, Berkeley, respectively, where they continue their investigation of optical properties of quantum dots. Before 1993, QDs were prepared in aqueous solution with added stabilizing agents to avoid colloid precipitation (see colloid stability in Chapter 7, p. 518 "Intermolecular Forces"). In 1993, Bawendi and coworkers synthesized better luminescent CdSe QDs by using a high-temperature organometallic procedure instead.[10,11] Crystallites from 12 to –115 Å in diameter with consistent crystal structure, surface derivatization, and a high degree of monodispersity were prepared in a single reaction based on the pyrolysis of organometallic reagents by injection into a hot coordinating solvent. At that time, even if one could make QDs in a narrow size range, they still came with two major problems: 1) poor fluorescence and 2) hydrophobicity, making them useless in biology. The addition of semiconductor caps, such as ZnS or CdS caps over a CdSe core, was found to dramatically increase the fluorescence quantum yield to 45% or higher.[12–14] The core determines the nanocrystal color, and the shell of the higher bandgap material (ZnS or CdS) dramatically enhances not only the brightness but also the chemical stability. A few different methods for making nanocrystals water soluble are also available today.[15] Methods for water solubilization of QDs include derivatizing the surface with mercaptoacetic acid or dithiothreitol, and Alivisatos and coworkers use a silica/siloxane coating.[16,17] The addition of these coats makes the particles water soluble, and therefore more useful in biology experiments.

The quest to find new energy solutions stimulated the development of modern quantum dot technology even further. The advantage of the surface area-to-volume ratio of nanocrystal particles for energy conversion was realized, and photoelectrochemistry research (e.g., solar energy conversion) tapped the semiconductor/liquid interface to exploit this (also visit http://www.technologyreview.com/Energy/19256).[18,19]

Quantum dots, we will learn in Chapter 7 and Volume III, Chapter 3, are manufacturable through a very wide variety of methods, including self-assembly in colloidal wet chemical synthesis, template chemistry (zeolite, alumina templates), sol-gel methods, micelle methods, organometallic synthesis, pyrolysis, lithography and etching, electrostatic confinement, scanning tunneling microscope (STM) tips, epitaxial strain, and so on. QD devices have now been demonstrated in many research laboratories, and commercial products are available on the market.

Zero Point Energy We noted before that at the lowest electron energy (n = 1), the *ground state energy,* remains finite despite the fact that V = 0 inside the solid (see Equation 3.156 and Figure 3.58c). In classical physics, both the kinetic energy (KE) and potential energy (PE) can have a zero value, but not in quantum physics! According to quantum mechanics, an electron in a box of length L must necessarily have energy. It cannot be inside the box and have zero energy. The minimum possible energy is E_1 and is called the zero point energy (ZPE). Zero point energy hypothesized by Max Planck in 1911 and developed by him and Walter Nernst between 1911 and 1916. It was first measured by Dr. Willis Lamb in 1947 as a slight upward shift of electrons in their atomic orbitals. The concept of a zero point energy has very far-reaching consequences, manifesting

themselves from the quantum world to the cosmos. The existence of a zero point energy for a quantum particle is a general phenomenon and is consistent with the Heisenberg uncertainty principle; i.e., particles with zero momentum cannot be localized. From the Heisenberg uncertainty principle (Equation 3.106), it follows that for a 1D box:

$$E = \frac{p^2}{2m} \text{ and also}$$

$$\frac{(\Delta p)^2}{2m} \geq \frac{\hbar^2}{2m(\Delta x)^2} \geq \frac{\hbar^2}{2m\left(\frac{L}{2}\right)^2} = \frac{2\hbar^2}{mL^2} > 0 \quad (3.162)$$

where we used $\Delta x \leq \frac{L}{2}$, and also for E_1 (see Equation 3.156) we obtain:

$$E_1 = \frac{k_x^2 \hbar^2}{2m} = \frac{4\pi^2\hbar^2}{2mL^2} > \frac{2\hbar^2}{mL^2} > 0 \quad (3.163)$$

because $k_x = 2\pi/L$ (see Equation 3.161).

That the zero point energy is a direct consequence of the Heisenberg uncertainty principle can simply be recognized by considering an electronic vibration ceasing at T = 0. If such a thing could happen, then position and momentum both would be zero, which would violate the Heisenberg uncertainty principle.

The zero point energy can be measured through the Casimir force, a small attractive force between two close parallel, uncharged plates in a vacuum (see Volume III, Chapter 8 on actuators). This force, inversely proportional to the fourth power of the distance between the plates, comes about because, according to quantum physics, even a perfect vacuum contains a zero point energy keeping virtual particles in a continuous state of fluctuation (see Equation 3.107). Given the equivalence of mass and energy expressed by Einstein's $E = mc^2$, the vacuum energy must be able to create particles, known as virtual particles. They flash briefly into existence and expire within an interval dictated by the uncertainty principle. Casimir realized that only the virtual photons with wavelengths that fit a whole number of times into the gap between the two plates should be part of the energy calculation (i.e., no half-quantum lengths are allowed). Moving the plates together, the energy density should thus decrease as there are more allowed virtual photon states pushing against the plates from the outside than from between the two plates (see Figure 3.69).

An example of a practical application of the existence of the zero point energy for confined electrons is the quantum well laser, which is more efficient than a diode laser. In a quantum well, as we saw above, there are no allowed electron states at the very lowest energies, but there are more available states in the lowest conduction band state and at the top of the valence band so that many more holes and electrons can combine and produce photons with identical energy for enhanced probability of stimulated emission (lasing; see also Chapter 5).

Potential Wells of Finite Depth and Size, Finite Barriers, Tunneling, and Interfaces

Finite-Depth Potential Wells A potential well is a potential energy function with a minimum. An infinitely deep potential well, as considered above, is an idealization. On the atomic scale there are no infinitely high and sharp barriers, and both $\psi(x)$ and $d\psi(x)/dx$ go to zero smoothly near a boundary. If the walls of an electron *prison* are not infinitely high, as sketched in Figure 3.56, but can be scaled by the particles inside as sketched in Figure 3.70, Equation 3.137 must be rewritten as:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} = (E - V_0)\psi(x) \quad (3.164)$$

with V_0 the potential barrier the particle must jump to get out of the well. We will now solve this equation in the three regions marked in Figure 3.70.

A quantum well with finite potential walls for trapped particles is a more realistic picture than the infinite potential wells assumed earlier and applies

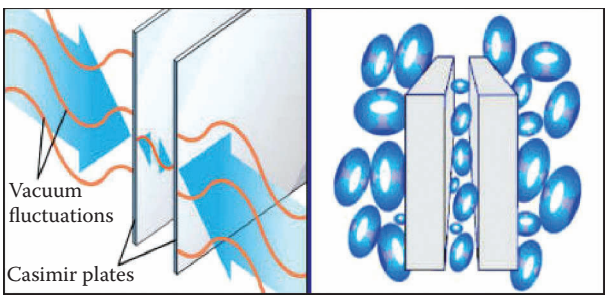


FIGURE 3.69 Vacuum fluctuations. The Casimir effect.

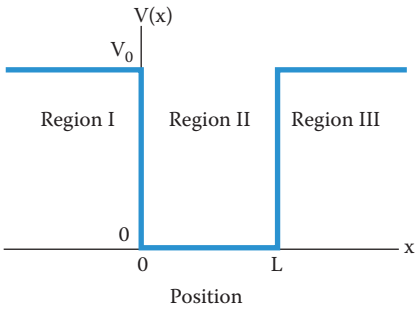


FIGURE 3.70 Quantum well with finite walls.

to most practical 3D (quantum dots), 2D (quantum wires), and 1D (quantum wells) confined structures. The representation also applies to nucleons inside the nucleus of an atom. Even the Krönig-Penney model, used to calculate the potential energy of an electron in a row of atoms in a linear solid, uses an array of periodic square wells of the type shown in Figure 3.70. In this model, each cell with V(x) = 0 represents an atom.

Outside the box sketched in Figure 3.70, in regions I and III, the boundary condition is that $V(x) = V_0$. These are regions that are "forbidden" to classical particles with $E < V_0$. With $E < V_0$ a classical particle cannot penetrate a barrier region: think about a particle hitting a metal foil and only penetrating the foil if its initial energy is greater than the potential energy it would possess when embedded in the foil and where otherwise it will be reflected. Defining α as:

$$\alpha^2 = \frac{2m(V_0 - E)}{\hbar^2} \tag{3.165}$$

yields:

$$\frac{d^2\psi}{dx^2} = \alpha^2\psi \tag{3.166}$$

In a region with $E < V_0$ there is an immediate effect on the waveform for the particle because $\mathbf{k}_x$ is real under these conditions, and we can write:

$$\mathbf{k}_x = \alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar} \tag{3.167}$$

We find for the general solution of Equation 3.164 in regions I and III a wave function of the form $\psi(x) = Ae^{+\alpha x} + Be^{-\alpha x}$, i.e., a mixture of an increasing and a decreasing exponential function. With a barrier that is infinitely thick we can see that the increasing exponential must be ruled out as it conflicts with the Born interpretation because it would imply an infinite amplitude. Therefore, in a barrier region the wave function must simply be the decaying exponential $Be^{-\alpha x}$, the important point being that a particle may be found inside a classically forbidden region (regions I and III).

If the barrier thickness is not infinite, then the increasing exponential component in the wave function cannot be ruled out because the wave function amplitude may not necessarily rise to infinity before the potential drops back to zero. In this case, the solutions for regions I and III are the following damped exponentials:

$$\psi_I(x) = Ae^{+\alpha x} \text{ (region I, } x < 0) \tag{3.168}$$

$$\psi_{III}(x) = Be^{-\alpha x} \text{ (region III, } x > L) \tag{3.169}$$

The values for *A* and *B* of the damped exponentials in Equations 3.168 and 3.169 are determined from the boundary conditions. These boundary conditions include the fact that wave function ψ and its derivative $d\psi/dx$ must be continuous at boundaries between regions I, II, and III (the boundary conditions require that $\psi_I = \psi_{II}$ at $x = 0$ and $\psi_{II} = \psi_{III}$ at $x = L$).

In Figure 3.71, the "leaky" waves in the forbidden regions are shown as exponential tails. The penetration depth of these waves is the distance outside the potential well over which the probability significantly decreases and is given by:

$$\delta x \approx \frac{1}{\alpha} = \frac{\hbar}{\sqrt{2m(V_0 - E)}} \tag{3.170}$$

Thus, the penetration distance that violates classical physics is proportional to Planck's constant and also depends on the value of $V_0 - E$ and on the mass of the particle. Because of this "barrier penetration," the electron density of a material extends outside the surface of the material.

In region II ($0 < x < L$) where V(x) = 0, the wave equation becomes:

$$\frac{d^2\psi}{dx^2} = -\mathbf{k}_x^2\psi \tag{3.171}$$

where:

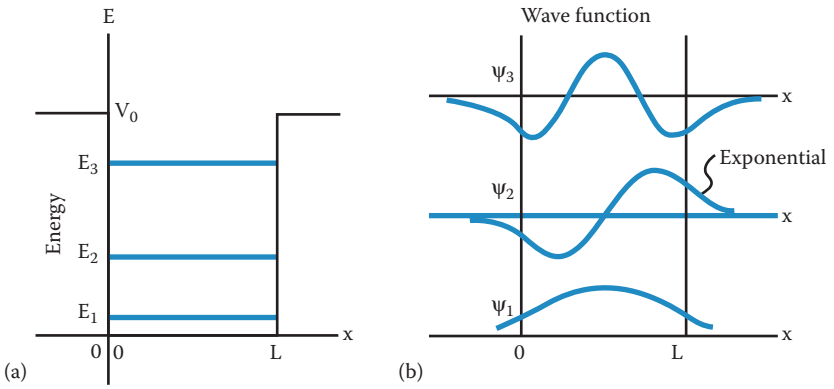


FIGURE 3.71 (a) Energies of a wave function inside the quantum well (region II). (b) The wave functions leak out of the well: see exponential tails.

$$\mathbf{k}_x = i\alpha \text{ and } \alpha = \frac{\sqrt{2mE}}{\hbar} \tag{3.172}$$

The solution here is an oscillating wave just as in the case of the well with infinite walls (Figure 3.56):

$$\psi_{II} = Ce^{i\mathbf{k}_x x} + De^{-i\mathbf{k}_x x}$$

or (3.173)

$$\psi_{II}(x) = \psi_0 \sin \mathbf{k}_x x + \psi_0' \cos \mathbf{k}_x x$$

Thus, the wave functions for a particle in a well with finite walls look very similar to the ones for the infinite square well … except that the particle now has a finite probability of "leaking out" of the well!

Finite Height Barriers (Step Functions)

(1) $E > V_0$ A potential barrier is the opposite of a potential well. It is a potential energy function with a maximum. For a barrier of finite height and thickness (Figure 3.72) and with $E > V_0$, we use again the TISE (Equation 3.137). For particles outside and above the barrier (regions I and III), V = 0 and we obtain:

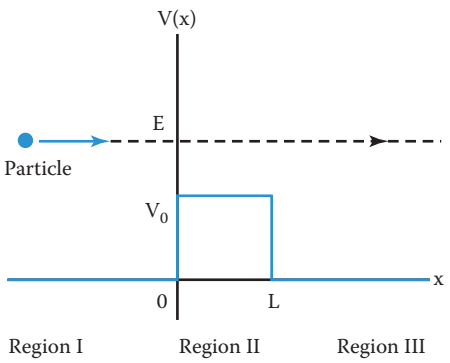


FIGURE 3.72 Finite barrier with boundary conditions.

$$\begin{aligned}\psi_I(x) &= Ae^{+ik_I x} + Be^{-ik_I x} (x < 0 \text{ and } V = 0)\\ \psi_{III}(x) &= Ee^{+ik_{III} x} + Fe^{-ik_{III} x} (x > L \text{ and } V = 0)\end{aligned} \tag{3.174}$$

These are oscillations with the wave vector:

$$\mathbf{k}_I = \mathbf{k}_{III} = \sqrt{\frac{2mE}{\hbar^2}} \tag{3.175}$$

We have replaced $\mathbf{k}_x$ here with the symbol $\mathbf{k}_I = \mathbf{k}_{III}$ or $\mathbf{k}_{II}$ to be more specific about the regime under consideration.

It is important to recognize here that with $E > V_0$ a particle would, in a classical picture, easily overcome the barrier, and one would expect a 100 % transmission. We will see that this is not the case when the particle is "wavy." In the barrier region (region II), where $0 < x < L$, we calculate:

$$\psi_{II}(x) = Ce^{+ik_{II} x} + De^{-ik_{II} x} \tag{3.176}$$

with the wave vector:

$$k_{II} = \frac{\sqrt{2m(E - V_0)}}{\hbar} \tag{3.177}$$

whereas in the case of $E < V_0$, $\mathbf{k}_x$ is real in a barrier region (see Equation 3.167), in the case of $E > V_0$, $\mathbf{k}_x$ (=$\mathbf{k}_{II}$) remains imaginary. If we are only considering waves moving from left to right (Figure 3.73), we can simplify the above wave functions to:

$$\text{Incident wave: } \psi_I(x) = Ae^{+ik_I x} \tag{3.178}$$

$$\text{Reflected wave: } \psi_I(x) = Be^{-ik_I x} \tag{3.179}$$

$$\text{Transmitted wave: } \psi_{III}(x) = Ee^{+ik_I x} \tag{3.180}$$

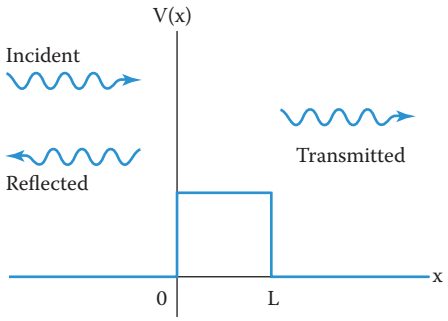


FIGURE 3.73 Wave with $E > V_0$ moving from left to right.

We define a "reflection coefficient" R as:

$$R = \frac{|B|^2}{|A|^2} \tag{3.181}$$

and likewise we can define a transmission coefficient T as:

$$T = \frac{|E|^2}{|A|^2} \tag{3.182}$$

The probability of the particles being reflected R or transmitted T is then:

$$R = \frac{|\psi_I(\text{reflected})|^2}{|\psi_I(\text{incident})|^2} = \frac{|B|^2}{|A|^2}$$

and:

$$T = \frac{|\psi_{III}(\text{transmitted})|^2}{|\psi_I(\text{incident})|^2} = \frac{|E|^2}{|A|^2} \tag{3.183}$$

The transmission probability is the probability that a particle incident on the left of the barrier emerges on the right of it.

(2) $E < V_0$ The situation where classically the particle does not have enough energy to surmount the potential barrier, $E < V_0$, is sketched in Figure 3.74.

With $E < V_0$, the particle will be reflected at x = 0 with the same kinetic energy and Equation 3.176 again applies, but the transmitted wave is now a damped exponential as we saw above. The wave function in region II becomes:

$$\psi_{II}(x) = \underset{\text{Unphysical} = 0}{Ce^{+\alpha x}} + \underset{\text{Damped}}{De^{-\alpha x}} \tag{3.184}$$

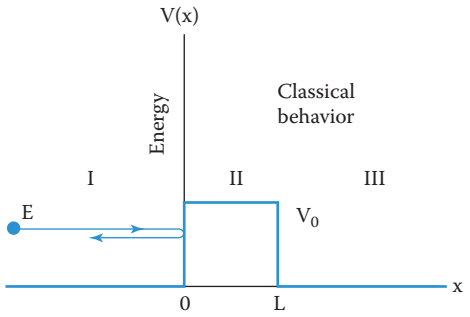


FIGURE 3.74 Wave with $E < V_0$ moving from left to right.

with $k_{II} = \alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar}$ (see Equation 3.167).

Tunneling We briefly return here to the tunneling phenomenon as described in Equation 3.167. The violation of classical physics in tunneling as described by this equation is allowed by the uncertainty principle. A particle can violate classical physics by ΔE for a short time, $\Delta t \sim \hbar/\Delta E$ (see Equation 3.107). The tunneling wave function is shown in Figure 3.75. The exponential decay of the wave function inside the barrier is given as:

$$\psi(x) = Ae^{-\alpha x} \tag{3.185}$$

with $\alpha^2 = \frac{2m(V_0 - E)}{\hbar}$ (Equation 3.165). If the barrier is narrow enough (L in Figure 3.75 is small), there will be a finite probability P of finding the particle on the other side of the barrier. The probability of an electron reaching across barrier L is:

$$p = |\psi(x)|^2 = A^2e^{-2\alpha L} \tag{3.186}$$

where A is a function of energy E and barrier height V_0. The probability of finding an electron

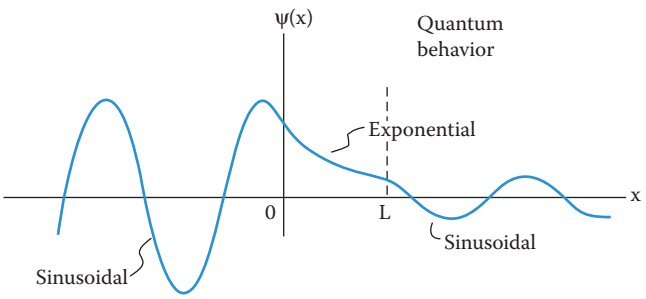


FIGURE 3.75 Tunneling wave function. The rectangular barrier stretches from x = 0 to x = L, the height of the barrier is V_0.

on the other side of a barrier of width L can be probed with a fine needle tip from a scanning tunneling microscope (STM) (see also Volume III, Chapter 6 on metrology). The tunneling current, based on Equation 3.186, picked up by the sharp needle point is given by:

$$I = f_w(E)A^2e^{-2\alpha L} \tag{3.187}$$

where $f_w(E)$ is the Fermi-Dirac function, which contains a weighted local density of electronic states in the solid surface that is being probed and states in the needle point (see also Fermi's golden rule further below). The weighted local density of electronic states is a material property of both probed surface and probe and may be obtained by measurements of the current I as a function of bias voltage V (dI/dV), which gives spatial and spectroscopic information about the quantum states of a nanostructure. From Equation 3.187, when L changes by 1 Å, the current changes by a factor of about 10! Obviously, the current is very sensitive to the gap distance. The size of the gap in practice is on the order of a couple of Angstroms! If the tip has two atoms vying for sitting at the very apex of the tungsten tip, the atom recessed by two atoms lower than the winning atom carries about 1 million times less current. That is why one wants such a fine tip.

Because particles must be either reflected (R) or transmitted (T) we have $R + T = 1$. By applying the boundary conditions $x \to \pm\infty$, $x = 0$, and $x = L$, we may also calculate the theoretical transmission or tunneling probability P from:

$$P = \frac{1}{1+G} \text{ with } G = \frac{(e^{\alpha L} - e^{-\alpha L})^2}{4\left(\frac{E}{V_0}\right)\left(1 - \frac{E}{V_0}\right)} \tag{3.188}$$

We note that P can be nonzero; i.e., particles may tunnel through a barrier even when $E < V_0$. The wave function does not fall abruptly to zero inside a region where its potential energy exceeds its total energy. This quantum mechanical result is a most remarkable feature of modern physics: there is a finite probability that the particle can penetrate the barrier and even emerge on the other side! We also notice here that even when $E > V_0$, $P < 1$. So a particle that has enough energy to overcome a barrier has a high probability to be reflected instead. This is counterintuitive as a classical particle with that energy would have a $P = 1$. We can summarize this situation as follows: quantum mechanics predicts an enhanced tunneling when $E < V_0$ and an enhanced reflection when $E > V_0$. This is equivalent to light reflecting from an interface with an abrupt change of refractive index.

Example 3.3: Assume that the work function (i.e., the energy difference between the most energetic conduction electrons and the potential barrier at the surface) of a certain metal is $\Phi = 5$ eV. Estimate the distance x outside the surface of the metal at which the electron probability density decreases to 1/1000 of that just inside the metal.

$$\frac{|\Psi(x)|^2}{|\Psi(0)|^2} = e^{-2kx} \approx \frac{1}{1000}$$

This leads to:

$$x = -\frac{1}{2k}\ln\left(\frac{1}{1000}\right) \approx 0.3 \text{ nm}$$

$$\text{Using } k = \sqrt{\frac{2m_e}{\hbar^2}(V_0 - E)} = 2\pi\sqrt{\frac{2m_e}{h^2}\Phi}$$

$$= 2\pi\sqrt{\frac{5\text{ eV}}{1.505\text{ eV}\cdot\text{nm}^2}} = 11.5\,\text{nm}^{-1}$$

with $m = m_e$.

In this section we considered the simplified problem of tunneling through a square barrier, but in most cases the barriers are not simply square shaped. One then needs to obtain a more general expression for the tunneling probability. These calculations are fairly involved, and we refer the reader to the specialized literature (e.g., Wolf's *Principles of Electron Tunneling Spectroscopy*[20]).

Some Tunneling History The concept of tunneling has no analogy in classical mechanics. The experimental manifestations of this effect are one of the many triumphs of the quantum theory. Based on electron tunneling, Fowler and Nordheim, in 1928, explained the main features of electron emission from cold metals by high external electric fields,

which had been unexplained since its first observation by Lilienfeld in 1922.

The discovery of the Esaki tunnel diode (see Figure 4.44) was very significant in the history of tunneling as it was the first electronic device where electron tunneling was clearly manifested in a semiconductor. Esaki described the tunnel diode in his 1957 thesis and received the 1973 Nobel Prize for his invention (http://nobelprize.org/nobel_prizes/physics/laureates/1973/esaki-bio.html). The next significant event was early in the 1970s, when the first quantum wells (QWs), which were also the first low-dimensional heterostructures, were demonstrated (see the AlGaAs/GaAs/AlGaAs structure in Figure 3.59).

Esaki was not only the originator of the Esaki tunnel diode; he also invented the double-barrier resonant tunneling diodes (abbreviated DBRT diode; see Figure 5.149) in 1974. Furthermore, in 1969, Esaki* and Tsu initiated research on semiconductor superlattices based on a periodic structure of alternating layers of semiconductor materials with wide and narrow bandgaps, in other words, a series of quantum wells or multiquantum well devices (MQWs).[21] The first superlattices were fabricated using an AlGaAs/GaAs material system (Figure 5.148).

In 1981, Gerd Binnig and Heinrich Rohrer invented the scanning tunneling microscope (STM), enabling the visualization and moving of individual atoms for the first time.

Tunneling is also of great importance in the history of nuclear physics. The decay of a nucleus is the escape of particles bound inside a barrier. The phenomenon of tunneling explains α-particle decay of heavy, radioactive nuclei. Inside the nucleus, an α-particle feels the strong, short-range attractive nuclear force as well as the repulsive Coulomb force. The nuclear force dominates inside the nuclear radius where the potential can be approximated by a square well. The rate for escape can be very small; particles in the nucleus "attempt to escape" 10^{20} times per second but may succeed in escaping only once in many years! Even if the quantum state (wave function) of the nucleus is completely defined with no uncertainty, one cannot predict when a nucleus will decay (Figure 3.76). Quantum mechanics tells us only the probability per unit time that any nucleus will decay (Figure 3.76a). Electrons, we saw above, are bound with negative total energy and can never escape the nucleus (Figure 3.76b). But a nucleon, such as an α-particle, is held inside the nucleus by the strong nuclear force and can escape by tunneling through the positive Coulomb barrier, which leads to the nuclear decay processes (G. Gamov, 1928). The potential barrier at the nuclear radius is several times greater than the energy of an α-particle. Based on Equation 3.107, for a short time, a particle can "borrow" energy from the uncertainty relation, gaining enough energy to jump over the potential barrier before giving it back. When it returns to its "proper" energy state, it is just outside the barrier instead of just inside, and rushes away. The process is as if the particle tunneled through the barrier, and it is purely a quantum effect.

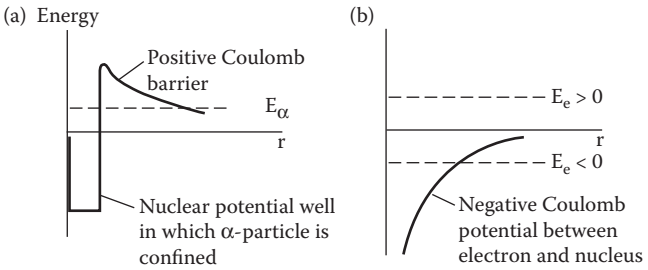


FIGURE 3.76 Nucleons escape the nucleus in radioactive decay (a); electrons cannot escape the nucleus (b).

Interfaces The theoretical treatment of particles hitting an interface is the same as the treatment of particles hitting one side of a square barrier (half square or half step function) and finds all types of applications. The results allow one to handle important problems involving the transmission and reflection of particle waves, such as encountered in vacuum/metal interfaces (work function), vacuum/semiconductor interfaces, semiconductor/metal interfaces, metal/metal interfaces (e.g., in thermocouples), and semiconductor/semiconductor interfaces (e.g., in diodes). For this analysis no new equations are needed as we can retrieve all the necessary expressions from our treatment of the full square barrier as summarized in Table 3.7.

* Esaki worked for IBM at the Thomas J. Watson Research Center, Yorktown Heights, New York, until 1992, when he returned to Japan to become president of Tsukuba University, Ibaraki.

TABLE 3.7 Transmitted and Reflected Particle Waves on an Interface

Energy	Equations	Wave Vectors
$E = \frac{\langle p^2 \rangle}{2m} > V_0$	Incident $\psi_I(x) = Ae^{+ik_I x} + Be^{-ik_I x}$ Reflected Transmitted $\psi_{II}(x) = Ce^{+ik_{II} x} + De^{-ik_{II} x}$ Unphysical = 0	$k_I = \frac{\sqrt{2mE}}{\hbar}$ $k_{II} = \frac{\sqrt{2m(E - V_0)}}{\hbar}$
$E = \frac{\langle p^2 \rangle}{2m} < V_0$	Incident $\psi_I(x) = Ae^{+ik_I x} + Be^{-ik_I x}$ Reflected Unphysical = 0 $\psi_{II}(x) = Ce^{+\alpha x} + De^{-\alpha x}$ Damped	$k_I = \frac{\sqrt{2mE}}{\hbar}$ $\alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar}$

Harmonic Potential Wells

We now revisit the harmonic oscillator from Figure 3.51 and solve it in the Schrödinger way so we may learn how quantization comes about in the case of an oscillating system. Simple harmonic oscillators describe many physical situations from springs to diatomic molecules and atomic lattices. In all cases, harmonic oscillations occur because the system contains a part that experiences a restoring force (spring) proportional to the displacement from equilibrium. In Figure 3.77, depicting a harmonic potential well, we discern two regions: inside the well (region I), with $E > V(x)$, and outside of the well (region II), with $E < V(x)$.

The time-independent Schrödinger equation (TISE) for a harmonic oscillator is given as:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} = \left(E - \frac{1}{2}kx^2\right)\psi(x) \qquad (3.189)$$

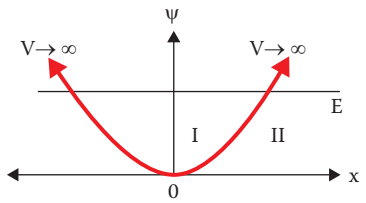


FIGURE 3.77 Harmonic potential well.

where $V(x) = \frac{1}{2}kx^2$ (in this equation k is the force constant, not the wave number **k**!) and $\omega = \left(\frac{k}{m}\right)^{\frac{1}{2}}$. The general solutions for Equation 3.189 are bell-shaped Gaussian functions multiplied by a Hermite polynomial $H_n(y)$:

$$\psi_n(x) = N_n H_n(\alpha^{\frac{1}{2}}x)e^{-\frac{\alpha x^2}{2}}$$

$$\text{with } n = 0,1,2,3\ldots \text{ and } \alpha = \left(\frac{K_m}{\hbar^2}\right)^{\frac{1}{2}}$$

$$N_n(\text{a normalization constant}) = \frac{1}{(2^n n!)^{\frac{1}{2}}}\left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} \qquad (3.190)$$

Some values for $H_n(y)$ are listed in Table 3.8.

The energy is given by:

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega, \; n = 0,1,2,3\ldots \qquad (3.191)$$

with n the vibrational quantum number. As shown in Figure 3.78, energies are evenly spaced and again cannot be zero in the ground state because of quantum confinement. Here we start the quantum numbers from n = 0 rather than n = 1.

The ground state (n = 0) of the wave inside the box is a simple Gaussian with no zero crossings (Figure 3.79a):

$$\psi_{n=0}(x) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} e^{-\frac{\alpha x^2}{2}} \qquad (3.192)$$

with zero point energy:

$$E_0 = \frac{1}{2}\hbar\omega \qquad (3.193)$$

TABLE 3.8 Hermite Polynomials, $H_n(y)$

n	$H_n(y)$
0	1
1	2y
2	$4y^2 - 2$
3	$8y^3 - 12y$
4	$16y^4 - 48y^2 + 12$
5	$32y^5 - 160y^3 + 120y$

n = vibrational quantum number.

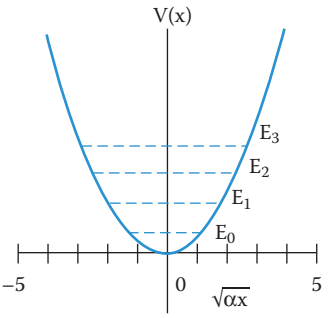


FIGURE 3.78 Energies of the quantum oscillator according to Equation 3.191, $\alpha = \left(\frac{K_m}{\hbar^2}\right)^{\frac{1}{2}}$.

The first excited state (n = 1): $\psi_{n=1}(x) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} \sqrt{2\alpha}xe^{\frac{-\alpha x^2}{2}}$ has one zero crossing; the second excited state (n = 2): $\psi_{n=2}(x) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} \frac{1}{\sqrt{2}}(2\alpha x^2 - 1)e^{\frac{-\alpha x^2}{2}}$ has two zero crossings; the third excited state (n = 3): $\psi_{n=3}(x) = \left(\frac{\alpha}{\pi}\right)^{\frac{1}{4}} \frac{1}{\sqrt{3}} \sqrt{\alpha}\, x(2\alpha x^2 - 3)e^{\frac{-\alpha x^2}{2}}$ has three crossings; and so on. From Figure 3.79b the lowest energy state (n = 0) of the QM oscillator has a maximum probability at the equilibrium position, whereas the classical oscillator always has its maximum probability at the extremes.

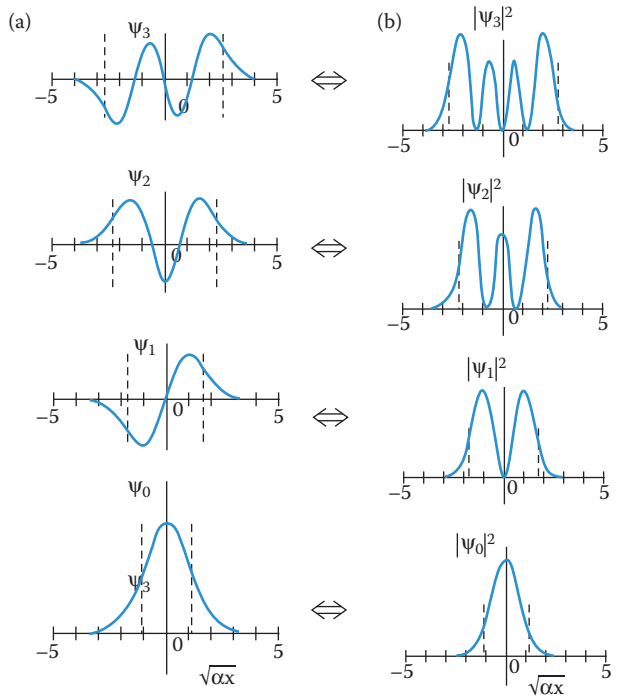


FIGURE 3.79 (a) The waves for n = 0 to n = 3. (b) Probabilities for the same waves.

As expected from Bohr's correspondence principle, the higher the quantum numbers, the better the quantized oscillator resembles the classical nonquantized oscillator from Figure 3.51. This is illustrated in Figure 3.80.

Quantized harmonic oscillators are all around us: diatomic molecules, vibrations within molecules, vibrations of atoms about equilibrium positions, oscillations of atoms or ions in crystal lattices (phonons), normal modes of electromagnetic fields in a cavity (blackbody radiation), Landau levels in the quantum Hall effect, and to a first approximation, any oscillatory behavior. We will encounter another example of the harmonic quantum oscillator solution presented here in the solution Einstein proposed in 1903 for the energies of the "atomic oscillators" in solids (see below under "Classical and Quantum Oscillators").

Example 3.4: The spacings between vibrational levels of molecules in the atmosphere, CO_2 and H_2O, are in the infrared frequency range: $\Delta E = h\nu = \hbar\omega \sim 0.01$ eV. As a consequence, Earth has an atmosphere acting as a greenhouse.

Particles in an Atom: Central Force

An electron bound to the hydrogen nucleus is an example of a central force system: the force depends on the radial distance between the electron and the nucleus only. The potential energy associated with a central or inverse square law force is very important

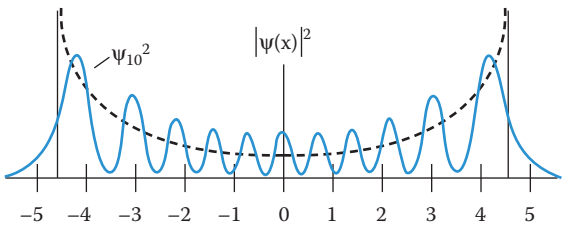


FIGURE 3.80 Quantized oscillator. Probability densities *P* for n = 10 states of a quantum mechanical harmonic oscillator. The probability densities for classical harmonic oscillators with the same energies are shown in black. In the n = 10 state, the wavelength is shortest at x = 0 and longest at x = |L|. The higher the quantum number *n*, the closer the solution resembles the classical one represented in Figure 3.51.

and was illustrated earlier for the case of a hydrogen atom in Figure 3.53. The potential energy of the electron-proton system is:

$$V(r) = -\frac{e^2}{4\pi\varepsilon_0 r} \tag{3.194}$$

The solutions of Schrödinger's equation with this potential are spherical Bessel functions. In Figure 3.53 the probability of finding the ground state for a hydrogen electron (n = 1) as a function of the radial distance from the proton was shown. The value of $|\Psi(x,t)|^2$ at some location at a given time is proportional to the probability of finding the particle at that location at that time. The wave functions for hydrogen are like vibrating strings or membranes, but the vibrations are in three dimensions and they are described by spherical Bessel functions. In Figure 3.81 we show solutions for higher quantum number cases (up to n = 4). In Figure 3.81A we provide a summary, and Figure 3.81B details some of these solutions for hydrogen. The lowest shells are spherical, 1s (n = 1, l = 0, m = 0), 2s (n = 2, l = 0, m = 0), 3s (n = 3, l = 0, m = 0), and 4s (n = 4, l = 0, m = 0), and hold two electrons each. The next shells are three 2p orbitals (n = 2, l = 1 and m = 1, 0, −1), which are dumbbell shaped and fit into the atom along perpendicular axes. A total of six electrons can fit. Level three (n = 3) has one s orbital and three p's—just like the ones at level two—and five three d's (with m = −2, −1, 0, +1, +2). The d orbitals are shaped for the most part like four-leaf clovers. At level four (n = 4) we encounter for the first time the 4f orbitals. They hold 14 electrons and look somewhat like onion blossoms. The electrons in the 4f orbitals are the furthest removed from the nucleus, are easily excited, and exhibit thousands of distinct energy states. They are associated with the many unusual optical, magnetic, and catalytic properties of the rare earth elements.

We come back to the mathematical formulation of the solutions of Schrödinger's equation for a central force when analyzing energy levels for quantum dots. These "artificial atoms" have energy quantization just as like atoms and molecules.

Summary: Most Important Periodic Potential Profiles and the Sommerfeld Model

Summarizing, the quantization for three of the most important potential profiles leads to the following mathematical solutions of Schrödinger's equation: for the central force, we obtain spherical Bessel functions; for an infinite square well potential, sines, cosines, and exponentials; and for an oscillator, Hermite polynomials. The quantized energy levels for each case are summarized in Figure 3.82 (only the 2D case is shown). Notice that the separation between consecutive energy levels increases with increasing *n* for an infinite square well potential (*b*) are evenly spaced for a harmonic well potential (*c*), and in the case of an inverse square law potential (*a*) the separation becomes closer with the larger *n*.

Sommerfeld's model assumes that the electrons in a metal experience a constant zero potential so

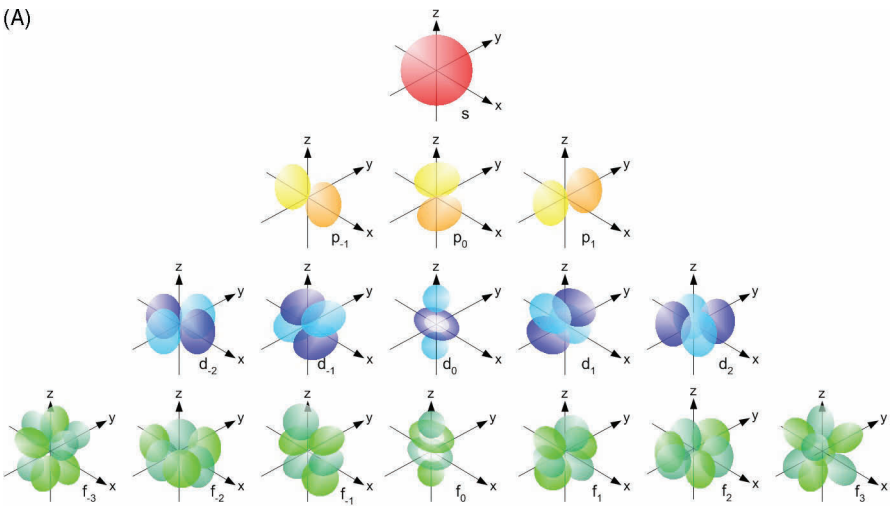


FIGURE 3.81 (A) Summary or overview of atomic orbitals up to n = 4.

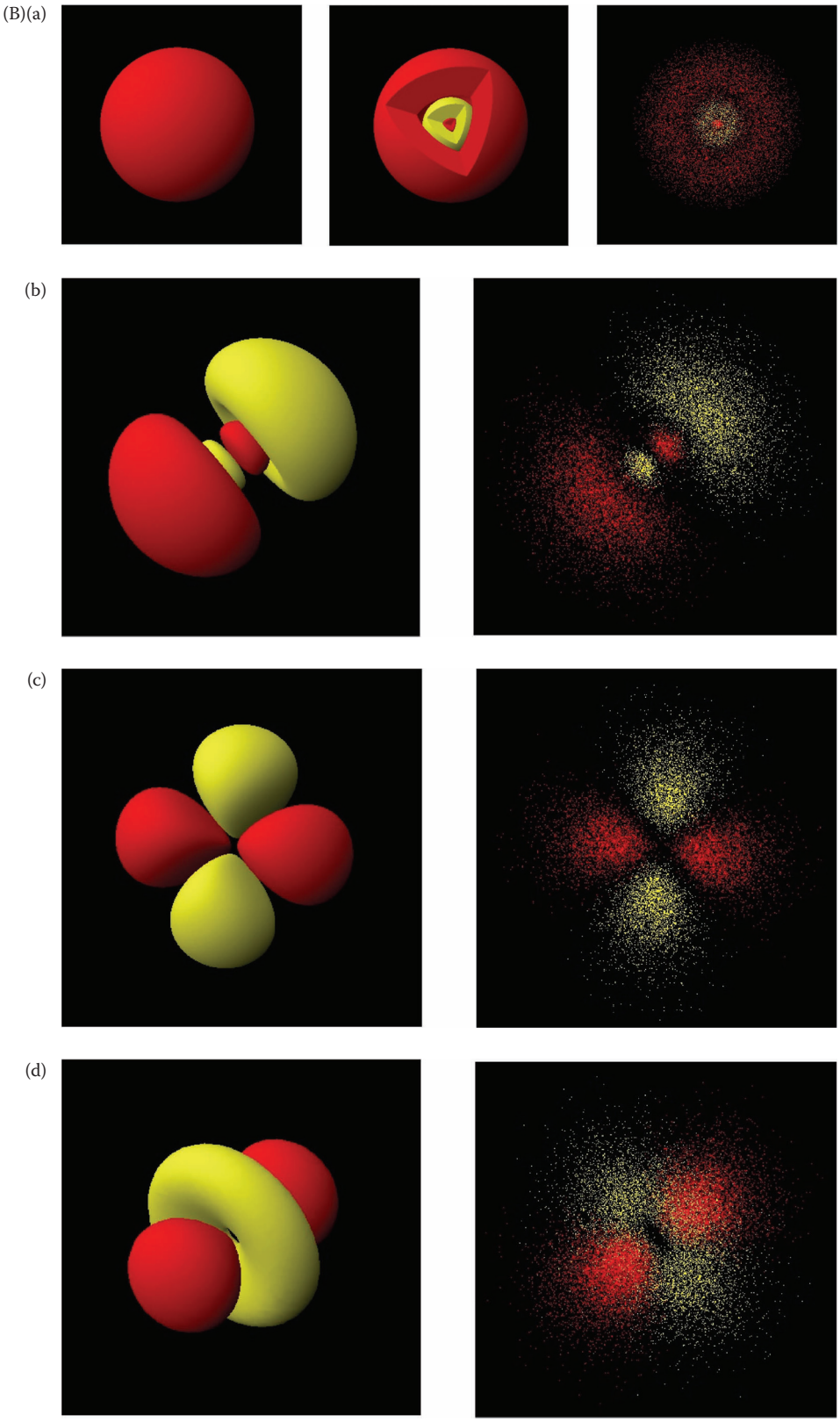


FIGURE 3.81 (*Continued*) (B) Where red = negative phase of Ψ and yellow = positive phase of Ψ. The density of dots reflects the magnitude of Ψ. (a) n = 3, l = 0, m = 0 for hydrogen. (b) n = 3, l = 1, m = –1 for hydrogen. (c) n = 3, l = 2, m = –1 for hydrogen. (d) 4 = 3, l = 3, m = 0 for hydrogen.

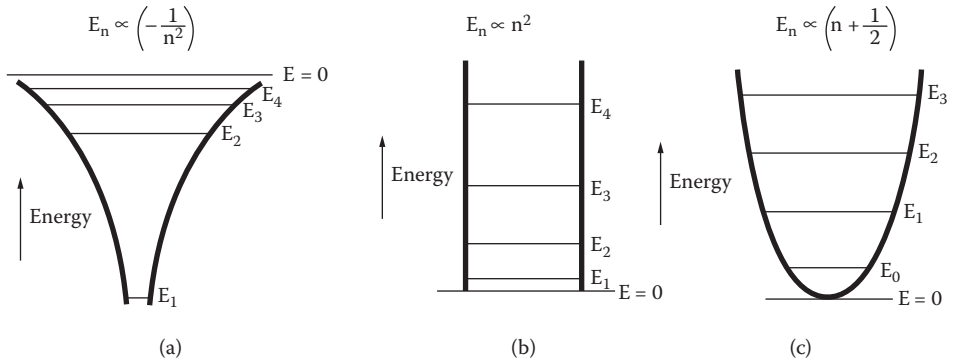


FIGURE 3.82 Quantized energy levels for a particle in an inverse square law potential (a) an infinite square well potential (b) and a harmonic well potential (c) 2D cases.

that the electrons are completely free to move about in the crystal. The important thing Sommerfeld's model does is introduce a finite surface for the metal at L and adopting a Fermi distribution for the charge carriers. What is most different from the classical theory as a result is the notion that only a few electrons, those with energies close to the Fermi level, contribute to the conduction mechanism. The Sommerfeld theory successfully explains specific heat, electrical conductivity, thermionic emission, thermal conductivity, and paramagnetism. However, the model fails to explain why some solids are good conductors, others are semiconductors, and yet others are insulators. The model also cannot account for the fact that some metals such as Be, Zn, and Cd exhibit a positive Hall constant; the free electron model always predicts a negative Hall coefficient! This model further predicts that the electrical conductivity is proportional to the electron concentration, but in reality divalent metals (Be, Cd, and Zn) and even trivalent metals (Al, In) have consistently lower conductivities than monovalent metals (Cu, Ag, and Au). Also, measurements of the Fermi surface, a concept introduced further below, indicate that it is often not spherical, contradicting Sommerfeld's model, which predicts a perfect sphere.

We prepare ourselves now to launch the more realistic Krönig-Penney model, where the potential energy of an electron in a row of atoms in a linear solid is modeled as an array of periodic square wells. We start by learning to combine atoms in simple molecules in the valence bond and molecular orbital theory; we then learn how to work with Bloch functions; and at last we are ready to introduce the band theory of solids.

Bringing Atoms Together

Molecules: Valence Bond and Molecular Orbital Theory

Molecules are formed from atoms by quantum mechanical forces. The so-called covalent bonds have no classical counterparts. They exist only because of the fermionic nature of the valence electrons. In the valence bond (VB) theory, atoms form electron-pair covalent bonds through the overlap of atomic orbitals of adjacent atoms. In the molecular orbital (MO) theory, shared, delocalized, valence electrons are viewed as occupying regions of space extending over all the binding atoms. When two atoms come together to form a molecule a valence bond is formed, and the valence electrons are in orbitals (called molecular orbitals) spread over the entire molecule, i.e., the electrons are delocalized. Half-filled atomic orbitals from the binding partners overlap and form bonds with two electrons of opposite spin occupying the molecular orbital. The latter corresponds to Pauli's exclusion principle, i.e., there are a maximum of two electrons with opposite spin per orbital. A diagram of the potential energy versus the internuclear distance for two approaching hydrogen atoms is shown in Figure 3.83a. The example involves the formation of a sigma bond (σ) from the overlap of s orbitals from the binding hydrogen atoms. At an internuclear distance, r_0, of 74 picometers, equilibrium is reached, and the attraction between the binding partners is maximized with an H-H bond strength of –436 kJ/mol. Hydrogen atoms react to form a molecule because the energy of the system is less than the sum of the individual constituents. From Figure 3.83b and c, we learn that compared

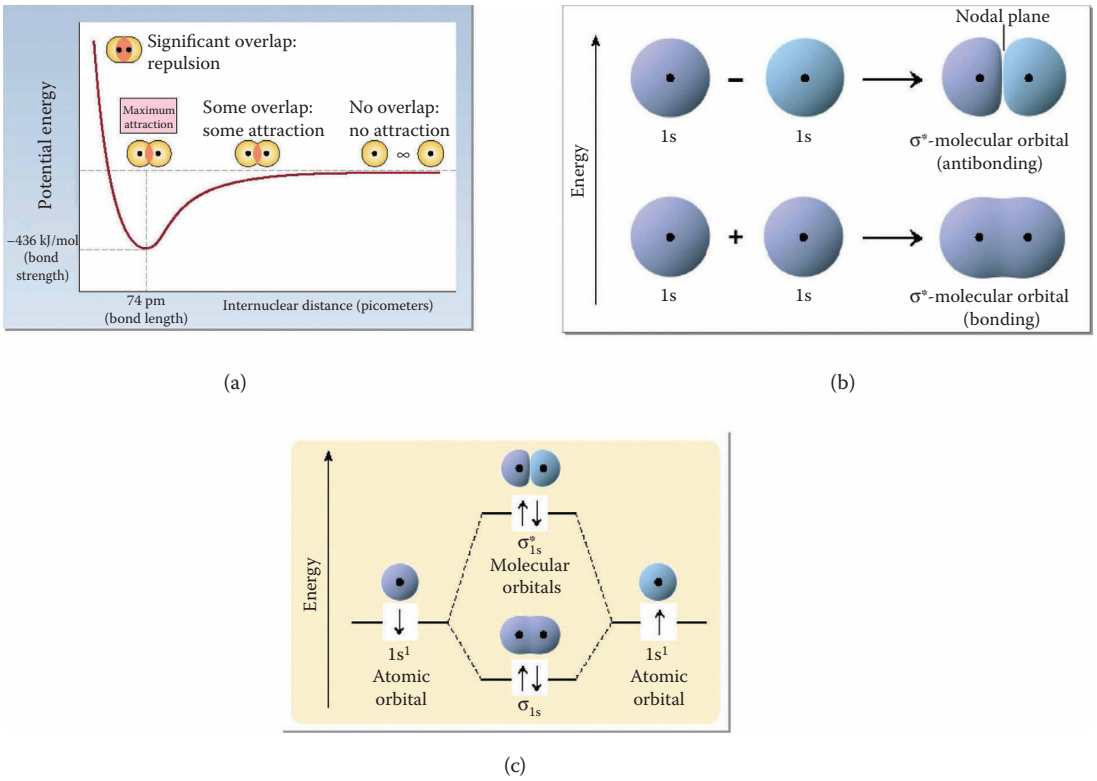


FIGURE 3.83 (a) A diagram of the potential energy versus the internuclear distance for two approaching hydrogen atoms. A sigma bond (σ) is formed from the overlap of the *s* orbitals from the binding hydrogen atoms. At an internuclear distance, r_0, of 74 picometers, equilibrium is reached, and the attraction between the binding partners is maximized with a H-H bond strength of –436 kJ/mol. (b) A bonding and antibonding molecular orbital is formed. (c) The bonding molecular orbital (σ_{1s}) is lower in energy than the parents' atomic orbitals, and the antibonding (σ^*_{1s}) is higher in energy.

with a single hydrogen atom there are twice as many theoretically permitted electron shells; the bonding molecular orbital (σ_{1s}) is lower in energy than the parents' atomic orbitals, and the antibonding (σ^*_{1s}) is higher in energy. The available electrons of the just created molecule are assigned, just as in the case of an atom, orbitals of successively higher energy and in the case of orbitals of equal energy—*degenerate orbitals*—these are filled one electron at a time before paring begins (Hund's rule), and only two electrons of opposite spin can occupy the same energy level (Pauli's exclusion principle). The total number of molecular orbitals equals the number of atomic orbitals contributed by the atoms constituting the molecule. When we examine a very large *N*-atom molecule, such as a long carbon chain hydrocarbon, we find a splitting of each one-atom energy level into *N* energy levels, each one corresponding to a somewhat different electron shell form.

Molecules also have excited states originating from a variety of sources. Because they are not usually spherically symmetric, molecules can rotate around several of their axes. The rotational energy is again quantized and depends on the *l* quantum number. The atoms in the molecule also vibrate compared to each other around an equilibrium distance. This can be described well by a harmonic oscillator spectrum (Equation 3.191).

To form a solid, one keeps adding atoms in three dimensions, making a very large molecule, and that is exactly what we will do next. We will see that in solids, just like in molecules, higher energy states occur where they interact to form a higher band of allowed energies.

Solids: A First Look at The Band Model of Solids

For the total number *N* of atoms in a solid (~10^{23} cm^{-3}), *N* energy levels split apart within a width ΔE at r_0 as shown in Figure 3.84a, where first two, then six, and then *N*, 3s atomic levels combine. This leads to a band of energies for each initial atomic energy level, and 2*N* electrons may occupy an energy band

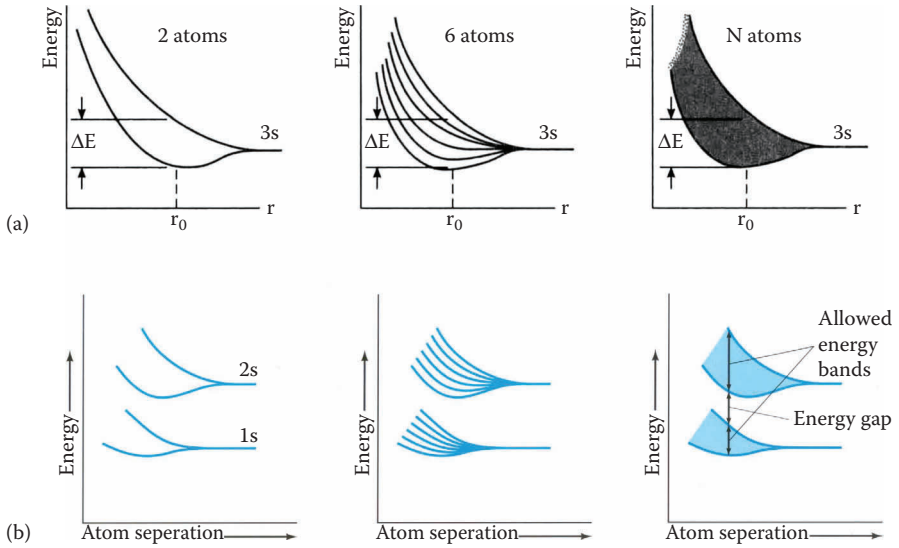


FIGURE 3.84 (a) Two, six, and then *N*, 3s atomic levels combine. This leads to a band of energies for each initial atomic energy level, and 2*N* electrons may occupy an energy band containing *N* energy levels. (b) Sets of 1s and 2s levels are combined into allowed energy bands separated by a forbidden energy zone, the bandgap, E_g.

containing *N* energy levels. In Figure 3.84b, sets of 1s and 2s levels are combined into allowed energy bands separated by a forbidden energy zone, the bandgap, E_g. A group of energy states incompletely filled at room temperature and empty at 0 K is called a conduction band. A conduction band is able to support the movement of electrons. A band of energy levels, missing some electrons at room temperature and completely filled at 0 K, is called a valence band. The valence band supports the movement of missing electrons, which are called holes. As illustrated in Figure 3.85, the relative position of conduction band, valence band, and forbidden bandgap are instructive for the classification of materials into conductors, semiconductors, and insulators. In conductors, the conduction band is partially filled, allowing electrons to move freely, or the valence band overlaps with the conduction band, again enabling free electron movement. In insulators there is a substantial forbidden energy gap between completely filled valence band and empty conduction band (~9 eV in Figure 3.85). Semiconductors are similar to insulators, but the bandgap is narrower, and electrons and/or holes are available at room temperature. Because in semiconductors the energy gap is small, thermal energy might suffice for some electrons to jump from the valence band to the conduction band; however, more often doping with impurities is needed to generate enough charge carriers at room temperature.

Altering the band structure, confining the geometry or the potential of electrons, leads to engineered states with very interesting unseen properties.

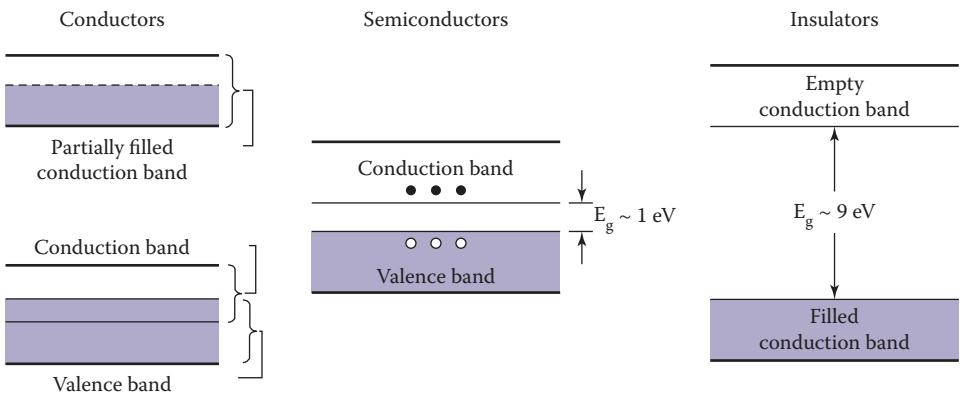


FIGURE 3.85 Classification of materials into conductors, semiconductors, and insulators based on the band model.

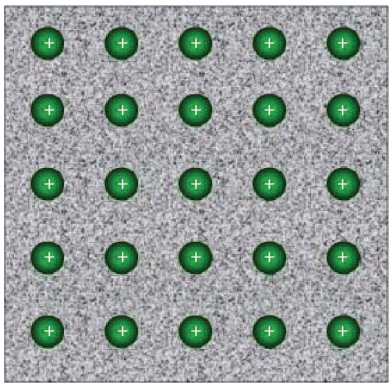


FIGURE 3.86 Positive ion cores in a metal.

Bloch Functions

Introduction Armed with some understanding of quantum mechanics and an introduction to the valence bond (VB) and molecular orbital (MO) theory, we are in a much better position to understand the band model of solids. The Drude free electron gas model surveyed above explained a number of important metallic properties, but it did not explain the most intriguing problems: what is it that distinguishes a metal from an insulator, and how can the same material, say carbon, form a conductor, a superconductor, a semiconductor, or an insulator? These facts remain unanswerable until one also takes into account that an electron gas moves through space occupied by a periodic array of positive-charged atomic cores (Figure 3.86). When electrons are free to roam in an infinite solid, e.g., in a metal, the electronic "orbitals" are traveling wave solutions, but when the metal is limited to a cube of size L^3 or a periodic array of positive ions is assumed, this will act to restrict the allowed energies and hence other quantities such as momentum, spin, and so on. Let us specify the force acting on the "free" electrons as $V(\mathbf{r})$.

Because the ion cores in a crystal are arranged periodically, the potential an electron feels is also periodic; with the periodicity of the underlying Bravais lattice in three dimensions, this yields the Born and von Karman's periodic boundary condition:

$$V(\mathbf{r}+\mathbf{R}) = V(\mathbf{r}) \tag{3.195}$$

for all Bravais lattice vectors $\mathbf{R}$ (see also Chapter 2). In Figure 3.87 we illustrate how the potential $V(\mathbf{r})$ for a single atom compares with that of two and with that of an array of atoms. Electrons in isolated atoms

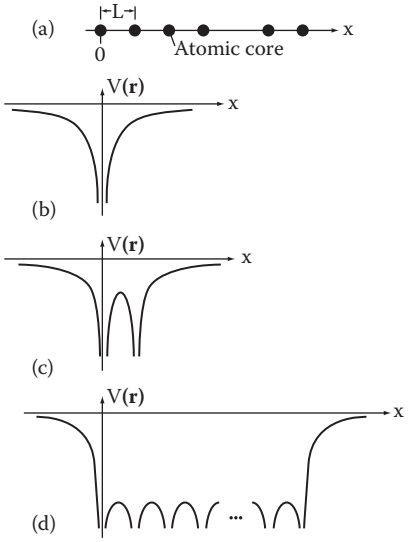


FIGURE 3.87 Bringing atoms together in a lattice. (a) Atomic cores; (b) isolated atom with binding energy as a function of $V(\mathbf{r})$; (c) two atoms brought together; and (d) array of atoms with binding energy as a function of $V(\mathbf{r})$.

occupy discrete allowed energy levels E_1, E_2, and so on, with a potential energy of the electron at a distance r from a positively charged nucleus q given as (see Figure 3.87a):

$$V(\mathbf{r}) = \frac{-qe}{4\pi\varepsilon_0 \mathbf{r}} \tag{3.196}$$

The 1D potential energy of an electron caused by an array of nuclei of charge q separated by a distance $\mathbf{R}$ is:

$$V(\mathbf{r}) = \sum_n \frac{-qe}{4\pi\varepsilon_0 |\mathbf{r} - n\mathbf{R}|} \tag{3.197}$$

where n = 0, ±1, ±2, and so on. This is shown in Figure 3.87d. The periodic potential in one dimension, shown here, is a Krönig and Penney-like potential (see Figure 3.98). From this figure, $V(\mathbf{r})$ is lower in the solid than in the isolated atoms. In the lowest binding energy states, conduction electrons move in a nearly constant potential as shown in Figure 3.88.

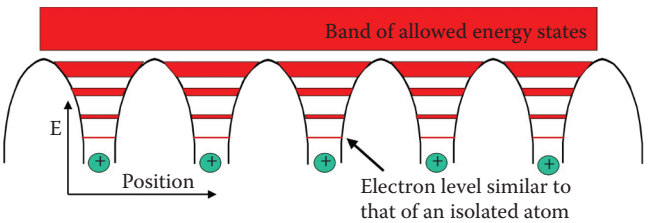


FIGURE 3.88 Band formation.

These electrons are trapped within the metal; a work function of several electron volts prevents them from escaping from the surface.

The higher energy states of tightly bound electrons are very similar to those of the isolated atoms. Lower binding electron states from the atoms become bands of allowed states in the crystal, and we will learn shortly that with partially filled bands the solid becomes a conductor.

Consider now what potentials an electron would see as it moves through a crystal lattice (limited to 1D for now). The electrostatic potential, $V(x)$, is periodic such that $V(x + L) = V(x)$. Bloch's theorem states that because the potential repeats every L length, the magnitude of the wave function (but not necessarily the phase) must also repeat every L length. This is the case because the probability of finding an electron at a given point in the crystal must be the same as found in the same location in any other unit cell. Generalizing, according to Bloch's theorem (1928), if $\psi_0(\mathbf{r})$ is a solution of Schrödinger's equation in free space, then a solution in a potential field that is periodic with period $\mathbf{R}$ is a product of $\psi_0(\mathbf{r})$ and another function $V(\mathbf{r})$, which is itself periodic with period $\mathbf{R}$. This theorem is one of the most important formal results in all of solid-state physics because it tells us the mathematical form of an electron wave function in the presence of a periodic potential energy. Thus, the wave function of the electron in a periodic potential $V(\mathbf{r})$ has the form:

$$\psi(\mathbf{r}) = \psi_0(\mathbf{r}) \cdot V(\mathbf{r}) \tag{3.198}$$

These are the so-called Bloch functions. In the Bloch approach an electron is considered to belong to the crystal as a whole rather than a particular atom. We have encountered the solution for $\psi_0(\mathbf{r})$ already. It is a running wave that can be written as a sine or cosine function or, more generally, in the form $e^{i\mathbf{k}\cdot\mathbf{r}}$, where $\mathbf{k}$ is equal to $2\pi/\lambda$, λ being the wavelength of the electron in this case. The running wave represents the behavior of the free electron. In the presence of the periodic potential, the running wave is modulated to give:

$$\psi(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} \cdot V(\mathbf{r}) \tag{3.199}$$

where $V(\mathbf{r}+\mathbf{R}) = V(\mathbf{r})$ for all R in the Bravais lattice. The Bloch function can be interpreted (approximately) as describing the electron distribution within a single unit cell, and an overall phase variation term $e^{i\mathbf{k}\cdot\mathbf{r}}$ represents a phase difference of the wave function in adjacent unit cells. The latter can take on several values depending on the wave vector $\mathbf{k}$. For core electrons (those tightly bound to the nucleus) Equation 3.199 represents a strongly localized wave function similar to the electron orbitals around a hydrogen atom. The less strongly bound valence electrons are described by more extended wave functions that have significant amplitude between neighboring atoms. Free electrons are described by wave functions with a high energy E, such that the wave vector remains real between atoms [where $V(\mathbf{r})$ is high]. This important result shows that the wave function for the electron itself has the periodicity of the lattice. As a consequence, just like with x-rays or vibrations of atoms in crystals (phonons), only certain wavelengths (that is, energies) are permissible for electrons in crystals. Importantly, we shall see that this leads, among other things, to a natural distinction among metals, insulators, and semiconductors. From Equation 3.199 we must have standing waves in the crystal that have a period equal to a multiple of the period of the crystal's electrostatic potential (similar to a multilayer antireflection coating in optics). Indeed, when an electron travels in a solid and enters a region with a lower potential energy (e.g., closer to a positively charged atom), the kinetic energy goes up, and the wave function acquires a shorter wavelength. This behavior of electron waves entering a low potential region resembles that of light entering a high refractive index region. Just as for light, changes in wavelength give rise to reflections, in this case electron wave reflection. Inside a crystal, electrons experience a periodic potential caused by the regularly spaced atomic cores in the crystal lattice, leading to multiple electron wave reflections and electron wave interference. The multiple reflections result in eigenmodes that are affected by the exact shape of the periodic potential $V(x)$. For an infinite crystal, the eigenfunctions describe a state with a well-defined energy and a corresponding spatial distribution of the electron throughout the entire crystal. In a simple linear lattice with lattice spacing a, we have $V(\mathbf{r} + \mathbf{R}) = V(\mathbf{r})$, as shown above.

It is important to note that because the wave function repeats each unit cell, we only have to consider what happens in one unit cell to describe the entire crystal. Thus, we can restrict ourselves to values of **k** such that $-\pi/a$ to $+\pi/a$ [implying $\mathbf{ka} \leq 1$ or $(2\pi/\lambda)\, a \leq 1$], or we can describe the electron behavior in a solid with wave vectors that lie in the first Brillouin zone.

Bloch Function Applied to a Six-Atom Linear Lattice To get a better appreciation about Bloch functions, let us consider an array of six atoms (N = 6), as illustrated in Figure 3.89. The Bloch function for an electron in this linear solid is $\psi_n(x) = V(x)e^{ik_n x}$, with a function $V(x)$ that depends on the electronic states involved (see Figure 3.90). The Born and von Karman's periodic boundary condition (Equation 3.195) can only be satisfied if the wave vector **k** has N possible values from $\mathbf{k} = \pi/L$ to $\mathbf{k} = \pi/a$, with L the total length of the six-atom crystal and a the lattice spacing. In other words, $\mathbf{k}_n = n\pi/L$, with $n = 1, 2, 3\ldots L/a$, or there are $N = L/a$ states. This is really nothing else than the electron in a box problem, except that the box is now divided in six compartments!

For N = 6, there are six different superpositions of the atomic states that form the crystal states as shown in Figure 3.91 (where we only consider the 1s combinations).

Bloch Function for Metals Let us now take an example metal and determine which states its electrons

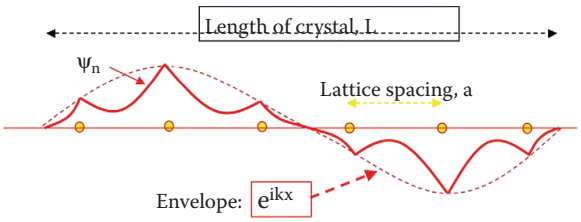


FIGURE 3.89 The states in a six-atom linear crystal (N = 6).

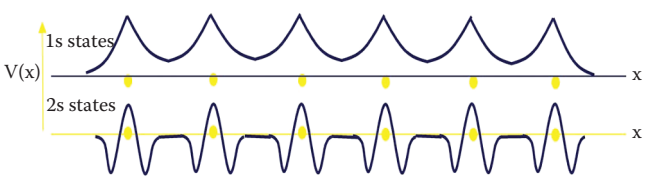


FIGURE 3.90 The periodic potential $V(x)$ depends on the electronic states of the atoms involved.

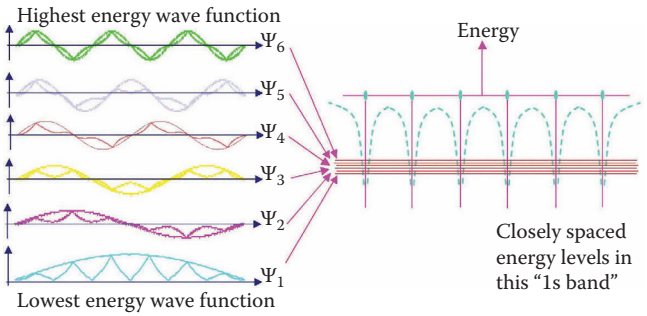


FIGURE 3.91 Six different super positions of the 1s atomic states form the crystal states.

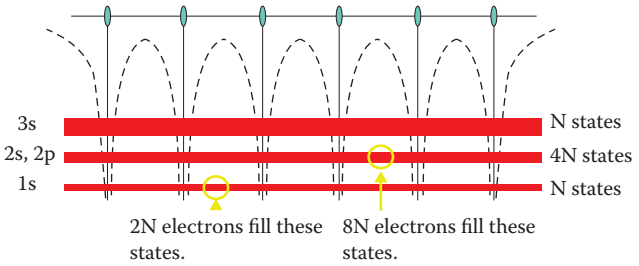


FIGURE 3.92 Filling the Bloch functions of sodium with electrons.

are occupying. In Figure 3.92 we have filled the Bloch functions of sodium with electrons according to Pauli's principle. We consider N sodium atoms, and because for sodium Z = 11 ($1s^2 2s^2 2p^6 3s^1$), this means we have $11N$ electrons to distribute. Notice that the 3s band is only half filled (N orbital states and N electrons). Electrons in this 3s band are easily promoted to higher states in the band (Figure 3.93), and this is what makes sodium a good conductor. In other words, to get a good conductor one needs a band partially filled with electrons.

The Bloch wave functions in metals are stationary waves, and in a perfectly periodic metal lattice, an electron would freely move without scattering from the atomic cores. This corresponds to a metal without any resistance at all! An electron in a periodic potential has a well-defined wave vector and momentum, and it is only when there are defects in the crystal,

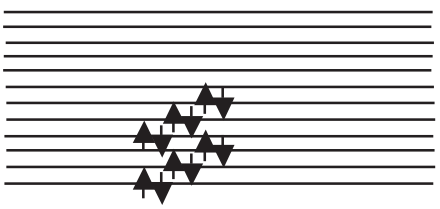


FIGURE 3.93 Half-filled 3*s* band of sodium.

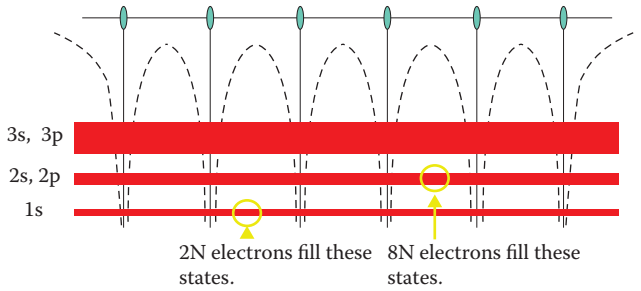


FIGURE 3.94 Filling the Bloch functions of silicon with electrons.

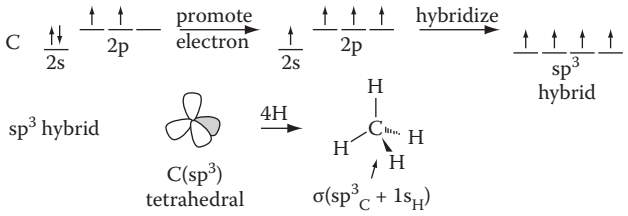


FIGURE 3.95 Hybridization of one 2s and three 2p orbitals in carbon to form four sp^3 hybrid orbitals. In methane carbon, four equivalent bonds are formed with hydrogen.

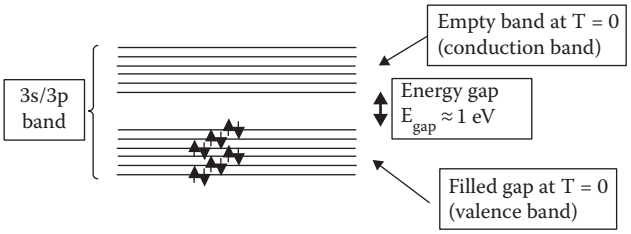


FIGURE 3.96 Two hybrid sp^3 bands in Si split in an empty conduction band and a filled valence band.

breaking up the lattice periodicity, that an electron may scatter to other Bloch states. Lattice vibrations also may break the perfect lattice periodicity, and as a consequence, electrons in metals scatter more at higher temperatures. This is all very different from Drude's hypothesis: it is not collisions with lattice atoms that determine the resistance of a metal but rather defects and lattice vibrations.

Bloch Functions for Semiconductors and Insulators

Next we will fill the Bloch function of a semiconductor such as Si with electrons. This is illustrated in Figure 3.94. In the case of Si, Z = 14, and the atom orbitals are $1s^2 2s^2 2p^6 3s^2 3p^2$. With a total number of N atoms, we will have $14N$ electrons to accommodate. At first blush it appears that, like Na, Si will also have a half-filled band because the $3s3p$ band has $4N$ orbital states and $4N$ electrons. By this analysis, Si should be a good metal, just like Na. But something unique happens for C and Si and other group IV elements.

In group IV elements, bonding orbitals are hybrid combinations of s and p states, so-called sp^3 hybrids. Hybrid atomic orbitals were introduced to reconcile the discrepancy between what atomic orbital theory predicts and what is seen experimentally. Take carbon, for example; the electron configuration is $1s^2$, $2s^2$, $2p^2$, so one expects carbon, in say methane, to form two bonds via the two unpaired p electrons; in practice we know that carbon forms four equivalent bonds. The hybridization process, in the case of C in methane, is explained in Figure 3.95. A $2s$ electron is promoted, and one $2s$ orbital and three $2p$ orbitals form four equivalent sp^3 hybrid orbitals used in binding four hydrogens to make a tetrahedral-shaped methane molecule. Similarly, in a Si lattice, four sp^3 orbitals link all the Si atoms tetrahedrally together. In the Si case, the hybridizing orbitals are one $3s$ orbital and three $3p$ orbitals, and these four equivalent bonding orbitals are completely filled in the single-crystal Si with two electrons each.

The superposition of sp^3 bands between neighboring Si atoms results in a filled bonding band, the valence band, and an empty antibonding band, the conduction band, as illustrated in Figure 3.96.

The electrons in a filled band cannot contribute to conduction because within reasonable E fields they cannot be promoted to a higher kinetic energy. Therefore, at T = 0, Si is an insulator. At higher temperatures, however, electrons are thermally promoted into the conduction band. For a Si crystal at room temperature, the amount of energy an electron must gain to overcome the bandgap is about 40 times more than the average amount of thermal energy. As a consequence, in a semiconductor the number n of free electrons increases rapidly with T (much faster than the scattering time τ decreases), whereas for a metal scattering time, τ gets shorter with increasing T (Figure 3.97).

Thus, energy bands and the gaps between them determine the conductivity and other properties of solids. Insulators have a valence band that is full and a large energy gap (a few eV). Consequently, no states of higher energy are available for electrons to go to.

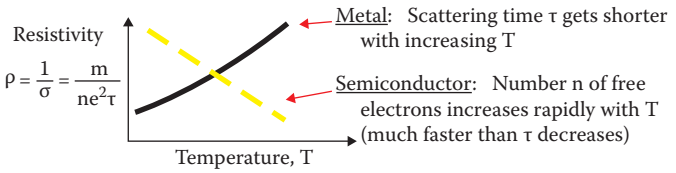


FIGURE 3.97 Resistivity as a function of temperature for a metal and a semiconductor.

Semiconductors are insulators at T = 0, but they have a small energy gap (~1 eV) between valence and conduction bands, so they conduct only at higher *T*. Metals have an upper band that is only partly full; in other words, the Fermi level lies within the valence band, and when applying an electric field, lots of states of higher energy are available for electrons to go to.

The Krönig-Penney Model

Solution of Schrödinger's Equation for a Periodic Potential

Bloch's theorem, along with the use of periodic boundary conditions, enables the calculation of the energy bands of electrons in a crystal if the potential energy function experienced by the electron is known. This was demonstrated for a simple finite square well potential model by Krönig and Penney in 1931, representing the first solution of Schrödinger's equation for a periodic potential. The Krönig-Penney model uses a simple 1D model of a crystalline solid as shown in Figure 3.98. The period of the potential is (a + b = L). The potential (V) is assumed equal to zero in the region of an atom, e.g., for $0 < x < a$, and to equal V_0 in the region between atoms, e.g., $-b < x < 0$. The calculations are a repeat of the calculations for a square barrier carried out above. For the zero regions, where electrons essentially act as free particles, the Schrödinger equation that applies is (Equation 3.171):

$$\frac{d^2\psi}{dx^2} = -k_x^2\psi$$

with (Equation 3.172):

$$k_x = i\frac{\sqrt{2mE}}{\hbar}$$

and the solution is an oscillating wave like in Equation 3.173:

$$\psi = Ae^{ik_x x} + Be^{-ik_x x}$$

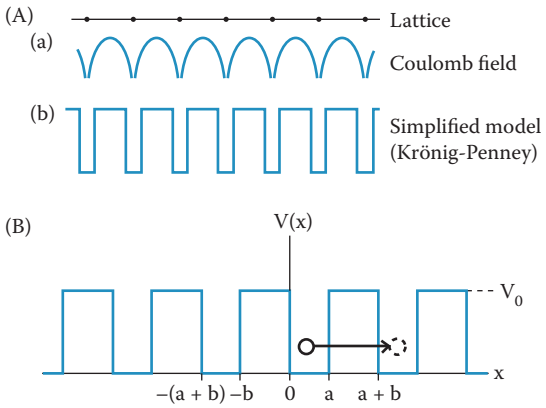


FIGURE 3.98 (A) (a) The periodic lattice potential in a real crystal (see also Figure 3.94). The bullets represent the positions of the nuclei. (b) One-dimensional periodic potential used in the Krönig-Penney model. A central question is whether an electron with energy *E* will be able to propagate from one lattice cell to another. (B) Electrons are essentially free between $0 < x < a$ (and in any similar region along the lattice) and have to tunnel through the barrier regions.

In the nonzero regions, electrons must tunnel through the rectangular barriers for which the following Schrödinger equation applies (Equation 3.166):

$$\frac{d^2\psi}{dx^2} = \alpha^2\psi$$

where α is (Equation 3.167):

$$\alpha = \frac{\sqrt{2m(V_0 - E)}}{\hbar}$$

and the solution is the summation of the damped exponentials of Equations 3.168 and 3.169:

$$\psi(x) = Ce^{+\alpha x} + De^{-\alpha x}$$

It is assumed that the energy E of the electron is always smaller than V_0; in other words, electrons stay in the solid.

We will not detail solving Schrödinger's equations for a linear array of atoms here. A detailed treatment can be found, for example, in A.J. Dekker[22] and better yet in S.O. Pillai.[23] Suffice it to say that one finds two solutions: one for the regions where V(x) = 0, and one for the regions where V(x) = V_0, and that these solutions come with constants *A*, *B*, *C*, and *D*. The boundary conditions are that the wave functions and their derivatives are continuous across the potential boundaries and that

the solutions must be Bloch functions of the form $\psi_n(x) = V(x)e^{ik_x x}$, with $V(x + nL) = V(x)$, $n = 0, +1, +2, +3, \ldots$, representing the symmetry of the assemblage of atoms. With application of these boundary conditions including the use of Bloch's theorem, the constants A, B, C, and D can de determined and the wave functions calculated. One can then show that, under the simplifying conditions, V_0 tends to infinity and b approaches zero, whereas the product $V_0 b$ remains finite, and solutions to the wave equations exist only if:

$$P\frac{\sin\beta a}{\beta a} + \cos\beta a = \cos ka \qquad (3.200)$$

where P is $\frac{mV_0 ba}{\hbar^2}$, a measure for the potential barrier height and width, and with β equal to $\left(\frac{2mE}{\hbar^2}\right)^{\frac{1}{2}}$. In Figure 3.99 we plot $P\frac{\sin\beta a}{\beta a} + \cos\beta a$, the left side of Equation 3.200, versus βa, fixing P at a value of $3\pi/2$, as a typical example. Because β^2 is proportional to the energy E, the x-axis in this figure is a measure of energy. Importantly, the right side in Equation 3.200, the term $\cos ka$, can only have values between –1 and +1, as marked by the two dashed horizontal lines in the figure. Obviously this condition can only be satisfied with values of βa for which the left side lies between +1 and –1.

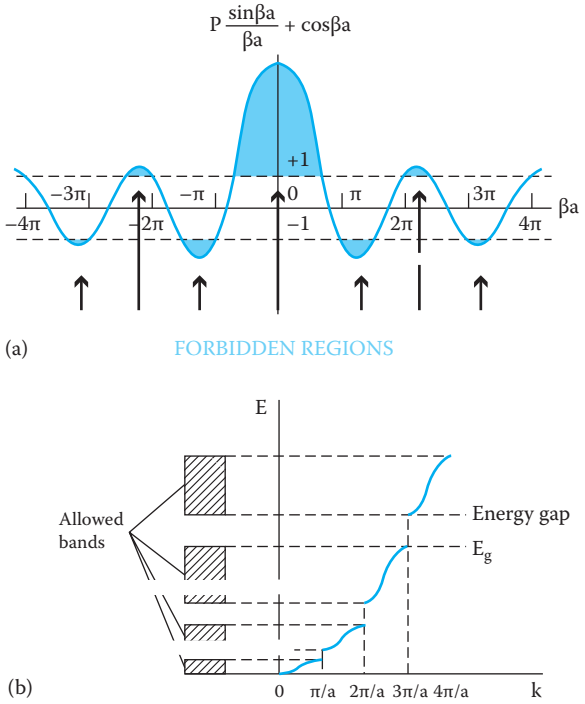


FIGURE 3.99 (a) Plot of $P\frac{\sin\beta a}{\beta a} + \cos\beta a$ versus βa. Arrows point to forbidden regions. (b) The energy as a function of wave number **k**. The allowed values of energy are given by those ranges of $\beta = \left(\frac{2mE}{\hbar^2}\right)^{\frac{1}{2}}$ for which the function lies between +1 and –1.

Analysis of the Solution

Some very important insights can be gained from an analysis of Equation 3.200 and an inspection of Figure 3.99:

1. Because β is related to E, electrons possess energies within certain bands but not outside those bands: there are allowed bands of energy and forbidden bands of energy (see arrows).
2. If V_0 increases, then P increases, the binding energy goes up, and the width of a particular allowed band decreases. The left side of Equation 3.200 becomes steeper, and in the limit, with P tending to infinity, the allowed energy bands reduce to the single energy levels we encounter in isolated atoms. In the latter situation, the equation only has solutions if $\sin\beta a = 0$, in other words, if $\beta a = \pm n\pi$ with $n = 1, 2, 3, \ldots$ and the energy spectrum becomes that of an electron in a constant potential box of atomic dimensions. In such cases, the energy levels are the energy levels of a potential well with $L = a$: $E_n = \frac{\pi^2\hbar^2}{2ma^2}n^2$ (see Equation 3.156). Each electron is confined to one atom by an infinite potential well, so electrons are completely bound to atoms.
3. In the case we decrease V_0 and reduce P to zero—in other words, when the binding energy goes to zero—Equation 3.200 is reduced to $\cos(\beta a) = \cos ka$ ($\beta = \mathbf{k}$), and the energy E is now given by $E = \frac{\hbar^2 k^2}{2m}$, i.e., the parabolic E-**k** relation for free electrons (see Figure 3.55). By varying V_0, the model thus covers the whole range from completely free electron to completely bound electron. In Figure 3.100 we illustrate the energy level structure as a function of varying degrees of binding strength.

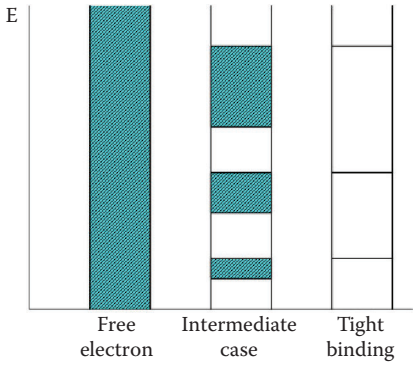


FIGURE 3.100 Energy level structure as a function of binding strength.

4. The width of allowed energy bands increases with increasing values of βa, i.e., with increasing energy, because the first term in the equation decreases on average with increasing βa.
5. From Figure 3.99a we see that at the boundary of an allowed energy band the cos (**k**a) = ±1 with **k** = nπ/a (dashed horizontal lines in the figure). Based on Equation 3.200 we can represent the energy also as a function of wave number **k**. The result is shown in Figure 3.99b and Figure 3.101 further below. This particular way of displaying the electronic levels in a periodic potential is known as the extended-zone scheme. Discontinuities in the *E*-**k** curve occur for **k** = nπ/a with n = 1, 2, 3,…. The zones in **k**-space that correspond to allowed energies for motion of an electron are the Brillouin zones (BZ), which we first encountered in Chapter 2.* The periodic potential $V(x)$ splits the free electron *E*-**k** curve into "energy bands" separated by gaps at each BZ boundary. Electrons can never have an energy within this energy gap. In a metal, the periodic potential $V(x)$ is very small or 0; in a semiconductor, the potential is large and consequently the energy splitting is large. The first Brillouin zone extends between **k** = –π/a and **k** = +π/a; the second is in the range of **k** from $-2\pi/a$ to $-\pi/a$ plus the range from π/a to $2\pi/a$. There can be no energy value for an electron between the bottom of the conduction band, E_c, and the top of the valence band, E_v. Therefore, the value $E_c - E_v = E_g$ is an energy gap at **k** = ±π/a. The existence of forbidden energies has very fundamental consequences. It happens for electrons in crystals with a periodically varying potential and for photons in systems with a periodically varying refractive index, in which case we call the gap a photonic bandgap (see Chapter 5). Brillouin zones are further detailed in the section below.

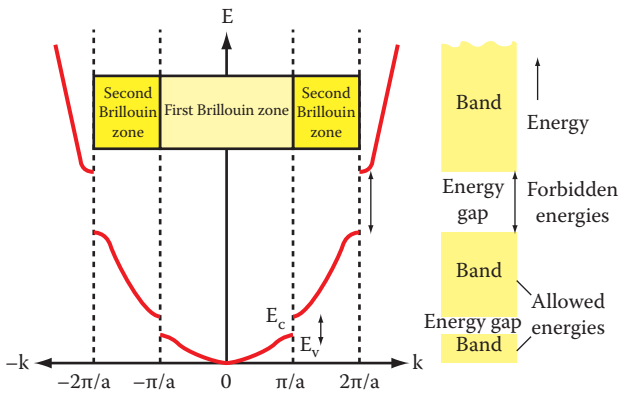


FIGURE 3.101 Energy versus wave number for motion of an electron in a one-dimensional periodic potential. The range of allowed **k** values goes from $-\pi/a$ to $+\pi/a$ corresponding to the first Brillouin zone for this system. Similarly, the second Brillouin zone consists of two parts: one extending from π/a to $2\pi/a$, and another part extending between $-\pi/a$ and $-2\pi/a$. This representation is called the extended zone scheme. Deviations from free electrons parabola are easily identified.

6. Finally, inspection of Equation 3.200 reveals that when **k** is substituted by $\mathbf{k} + 2\pi n/a$, where n is an integer, the right side of the equation remains the same, or **k** is not uniquely determined. In other words, in a given energy band, the energy is a periodic function of **k**. Given this periodicity, it is often convenient to introduce a "reduced wave vector" as $-\pi/a \le \mathbf{k} \le +\pi/a$. By shifting the second Brillouin zone $2\pi/a$ left or right, one can obtain the reduced zone scheme. A representation of energy versus reduced wave vector is marked by a double pointed arrow in Figure 3.102a. Obviously one does not need all *E*-**k** curves in all BZs. All information is already contained in the first Brillouin zone in a reduced zone scheme because of the $2\pi/a$ periodicity. In Figure 3.102a, we also display the electronic

* From Chapter 2 we remember that the x-ray diffraction pattern of a crystal is a map of the reciprocal lattice. It is a Fourier transform of the lattice in real space or also the representation of the lattice in **k**-space. This is true for the x-rays considered in Chapter 2 but also for the matter waves associated with electrons, neutrons, and so on. We are presenting BZs for all dimensionalities of the electronic structure further below.

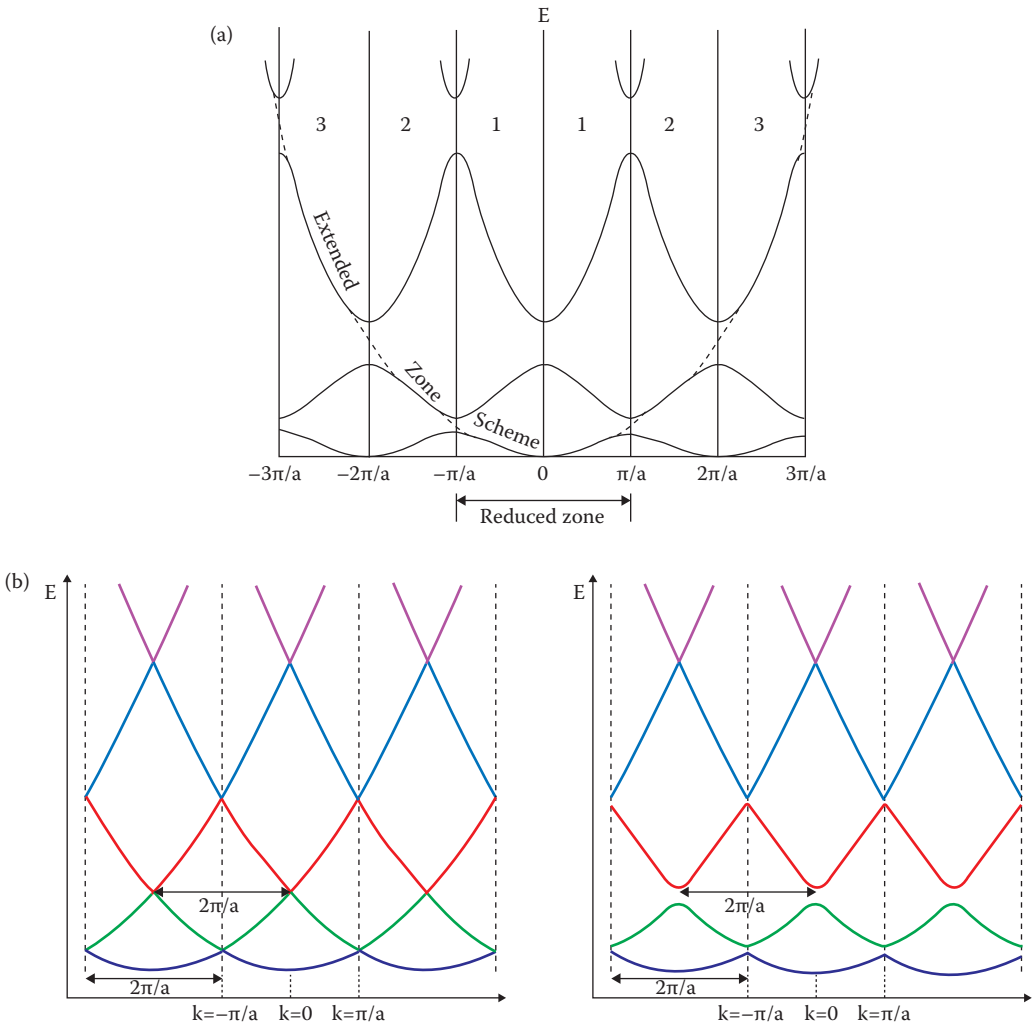


FIGURE 3.102 (a) Different representations of *E*-**k** in the presence of a periodic potential, band splitting. The extended zone scheme: plot *E*-**k** from **k** = 0 through all possible Brillouin zones (**bold curve**); the periodic or repeated zone scheme: redraw *E*-**k** in each zone and superimpose; the reduced zone scheme: all states with $|\mathbf{k}| > \pi/a$ are translated back into the first BZ. (b) Different *E*-**k** plots in the presence of a vanishing periodic potential: zone folding (left) and a nonvanishing periodic potential (right).

levels in a repeated or periodic zone scheme and repeat the extended zone scheme from Figure 3.101. In Figure 3.102b, we show what happens when the periodic potential becomes vanishing small; the bandgaps disappear and we get zone folding.

It is easy to show that the number of **k** values in each BZ is just *N*, the number of primitive unit cells in the sample. For this consider the finite, linear crystal with a total length of L = Na. The allowed values of the electron wave vector **k** in the first Brillouin zone for this arrangement are exactly N ($k = 0, \pm 2\pi/L, \pm 4\pi/L, \ldots, \pm N\pi/L$). Each primitive cell contributes exactly one independent value of **k** to each energy band. Thus, 2*N* electrons resulting from spin degeneracy can occupy each band. A monovalent element with one atom per primitive cell has only one valence electron per primitive cell and thus *N* electrons in the lowest energy band. This band will only be half-filled, and the material will be a conductor. The Fermi energy, the energy dividing the occupied and unoccupied states, for such a monovalent element will be in the middle of the valence band. If each atom contributes two valence electrons to the band (divalent element), the band, at T = 0 K, will be filled to the top. Thus, the simple rule whether an element is an insulator or a metal is given by whether the number of electrons is odd or even. This rule works surprisingly well;

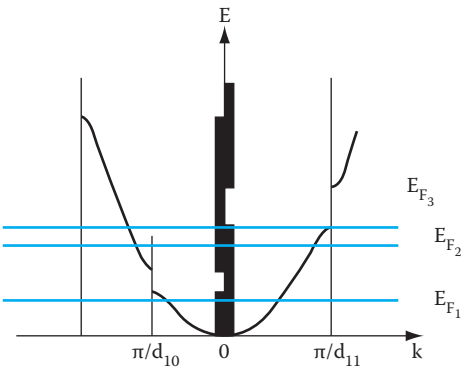


FIGURE 3.103 The presence of more than one periodicity in a crystal may cause the overlap in the integrated density of states as shown here as black rectangles along the *E*-axis. The material depicted in the *E*-**k** diagram is a semimetal conductor; see also Figure 3.104b.

the exceptions are caused by bands with more complicated structures in three dimensions than those discussed here. Indeed, the fact that a monovalent element is a conductor does not mean that a divalent element will always be an insulator. Although true in the 1D, it is not necessarily so in 2D or 3D! Bands along different directions in **k**-space can overlap, so that electrons can partially occupy both of the overlapping bands and thus form a semimetal. A semimetal is a metal with a carrier concentration several orders of magnitude smaller than the 10^{22} cm^{-3} typical for ordinary metals. Graphite and the pentavalent conducting elements, for example, are semimetals. A 2D band diagram for a semimetal is shown in Figure 3.103. In this example we show a material with three valence electrons and Fermi levels E_{F_1} to E_{F_3}. This case illustrates how bands in different directions overlap, resulting in a semimetal (this corresponds to the situation depicted in Figure 3.104b).

However, it remains true that only crystals with an even number of valence electrons in a primitive cell can be insulators. Depending on the energy band structure, for a given number of electrons, you can get a different filling, and it is the band structure that determines electronic and optical behavior. Some example band configurations leading to insulating and metallic behavior are shown in Figure 3.104.

The Effective Mass, Velocity of Charge Carriers, and Crystal Momentum

Up to this point we have implied that the mass of an electron in a solid is the same as the mass of a free electron (m_e). In reality, for some solids the measured electron mass is larger than that of the free electron, and in other cases it is smaller. The cause for this deviation is found in the interactions between the drifting electrons and the atoms in the crystal. The mass of an electron in a crystal is called the *effective mass*, m_e^*, and in general it is different from the mass of a free electron. The effective mass, m_e^*, of an electron is the mass of the free electron modified by the presence of the periodic potential of an array of positive lattice ions. Rather than moving undisturbed through the lattice, electrons are constantly jostled by atom movements (phonons). By lowering the temperature, atoms move more sluggishly, and this reduces the lattice resistance for electrons. However, temperature reduction does not reduce the

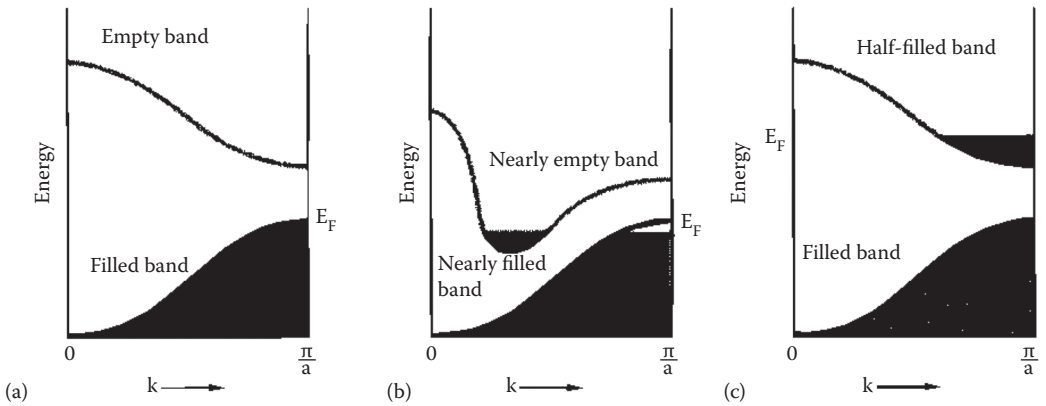


FIGURE 3.104 Occupied states and band structures leading to (a) semiconductor or insulator: two electrons per atom, *N* states per band, and two electrons/state; (b) semimetal: two electrons per atom, *N* states per band, and two electrons/state, a semimetal because bands overlap; (c) and metal: three electrons per atom (Li), *N* states per band, and two electrons/state, metal because of electron population.

influence of lattice defects; the only avenue here is to purify the crystal further, but even the purest crystal features some remaining defects. A third phenomenon controlling the speed of an electron through a crystal is the electron mobility, $\mu_e = \frac{e\tau}{m_e}$ (Equation 3.9). Applying the same voltage to equally pure samples of Si and GaAs, one finds that electrons in GaAs are accelerated much more. The electrons in Si and GaAs are of course the same, but their host lattices influence them differently. We refer to this property by saying that an electron has an effective mass in a given material; therefore, the effective mass m_e^* of an electron in gallium arsenide is less than that of an electron in silicon (making μ_e in GaAs larger; see Equation 3.9).

For a free particle such as an electron, we derived a wave solution $\psi(x) = Ae^{ik_x x}$ with energy $E = \frac{\hbar^2 k_x^2}{2m_e}\left(= \frac{p^2}{2m_e}\right)$ (Equation 3.144). For a free electron, the quantity $\hbar\mathbf{k}$ represents the true momentum $\mathbf{p}$ of the electron. For electrons in a crystal, we need to define a *crystal momentum* that is different from $\mathbf{p}$ because the energy for electrons in a crystal does not vary in the same fashion with $\mathbf{k}$ as it does for free electrons—in other words, the dispersion curve $E(\mathbf{k})$ is different for an electron in lattice than that of a free electron. When dealing with interactions of the electron with the lattice, we must use the conservation of crystal momentum $\hbar\mathbf{k}$ and not that of a free electron.

We will now show how the effective mass and the velocity of an electron in a lattice can both be derived from a knowledge of the energy dispersion curves $E(\mathbf{k})$. Remember that an electron state is a wave packet with a group velocity v_g given as the derivative of ω, the angular frequency of the de Broglie waves with respect to the wave number $\mathbf{k}$ or:

$$v_g = \frac{d\omega}{d\mathbf{k}} \tag{3.101}$$

From quantum theory the frequency of a wave function with energy E is $\omega = \frac{E}{\hbar}$ (because $E = \hbar\omega$), so that we may write:

$$\frac{dE}{d\mathbf{k}} = \hbar\frac{d\omega}{d\mathbf{k}} \tag{3.201}$$

Substituting this expression in Equation 3.101 results in:

$$v_g = \frac{1}{\hbar}\left[\frac{dE}{d\mathbf{k}}\right] \text{ or more generally } \mathbf{v}_g = \frac{1}{\hbar}\nabla_k E(\mathbf{k}) \tag{3.202}$$

The group velocity of a wave packet is basically the velocity at which the wave packet transports energy through the system. It also gives the average velocity of the Bloch electron. The influence the crystal is exerting on the electron motion is contained in the dispersion relation $E(\mathbf{k})$, and the velocity of an electron in a crystal then depends on the dispersion curve $E(\mathbf{k})$. Let us illustrate this point with a simple example. For a free electron, with the parabolic dispersion function $E(\mathbf{k}) = \frac{\hbar^2 k_x^2}{2m}$ (Equation 3.144), we derive $\frac{dE}{d\mathbf{k}} = \frac{\hbar^2\mathbf{k}}{m_e}$, and using Equation 3.202 we calculate for the free electron velocity:

$$v_g = \frac{\hbar\mathbf{k}}{m_e} = \left(\frac{\hbar\mathbf{k}}{m_e}\right)\left(\frac{2\pi\mathbf{p}}{h}\right) = \frac{\mathbf{p}}{m_e} \tag{3.203}$$

In this equation we have a linear relation of v_g with $\mathbf{k}$, which is of course expected because E is proportional to $\mathbf{k}^2$. In general, E is not proportional to $\mathbf{k}^2$, or at least only in small regions as schematically reproduced here in Figure 3.105d.

In Figure 3.105c, we show the velocity v as a function of $\mathbf{k}$. We observe that at the bottom of the energy band ($\mathbf{k} = 0$), the velocity is zero and then increases with $\mathbf{k}$ until it reaches a maximum value at $\mathbf{k} = \mathbf{k}_0$, corresponding to the inflection point in the E-$\mathbf{k}$ curve. Beyond the inflection point, the velocity starts to decrease and finally assumes the zero value at $\mathbf{k} = \pi/a$, which is at the top of the band.

An electron has a well-defined mass, and when accelerated it obeys Newtonian mechanics. To calculate the acceleration ($a = dv_g/dt$) of an electron in a crystal we derive from Equation 3.202 that:

$$a = \left(\frac{2\pi}{h}\right)\frac{d}{dt}\left(\frac{dE}{d\mathbf{k}}\right) \text{ or also } a = \left(\frac{2\pi}{h}\right)\frac{d^2E}{d\mathbf{k}^2}\frac{d\mathbf{k}}{dt} \tag{3.204}$$

The term $\frac{d^2E}{d\mathbf{k}^2}$ we can get from the E-$\mathbf{k}$ relationship, but we still need to derive $d\mathbf{k}/dt$ tocalculate a. For an

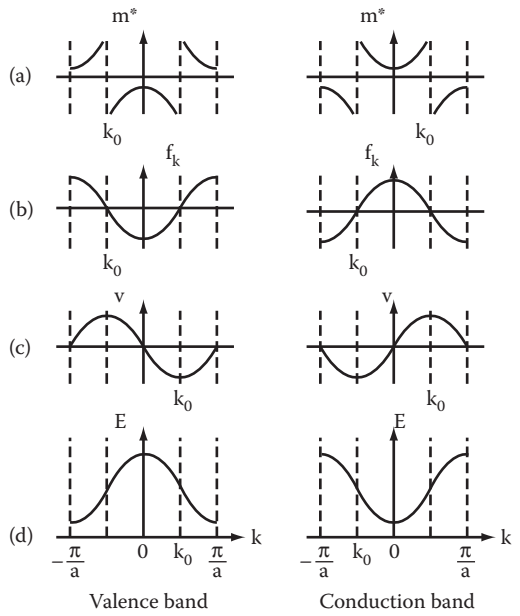


FIGURE 3.105 Effective mass m^*, f_k, velocity v, and energy E as a function of $\mathbf{k}_0$ for electrons in the conduction band (right) and holes in the valence band (left). See text for details.

electron with velocity $\mathbf{v}_g$, subjected to the influence of an external field E applied for a time dt, the distance traveled is $\mathbf{v}_g dt$, so that the work dE done by the electrical field on the electron is:

$$dE = -eE\mathbf{v}_g dt \tag{3.205}$$

Substituting the value for $\mathbf{v}_g$ from Equation 3.202 in the above equation, we get:

$$dE = -eE\left(\frac{2\pi}{h}\right)\left(\frac{dE}{d\mathbf{k}}\right)dt \tag{3.206}$$

or:

$$\frac{d\mathbf{k}}{dt} = -\frac{2\pi eE}{h} \tag{3.207}$$

or since $\mathbf{F} = -e\mathbf{E}$:

$$\hbar\frac{d\mathbf{k}}{dt} = \mathbf{F} \tag{3.208}$$

The last expression shows us that in a crystal $\hbar\dfrac{d\mathbf{k}}{dt}$ is equal to the external force on the electron, whereas in free space the force is equal to $d(mv)/dt$. The electron in a crystal is subject to both forces from the lattice and from external fields. In case there is also a magnetic field present, the force term $\mathbf{F}$ in Equation 3.208 must include the Lorentz force, $\mathbf{F}_L = -e(\mathbf{v}_g \times \mathbf{B})$ (see Equation 3.64), so that the equation of motion of an electron with group velocity $\mathbf{v}_g$ in a constant magnetic field $\mathbf{B}$ is given by:

$$\hbar\frac{d\mathbf{k}}{dt} = -e(\mathbf{v}_g \times \mathbf{B}) \tag{3.209}$$

Combining Equation 3.204 with Equation 3.207 (no magnetic field) we finally derive for a:

$$a = \left(\frac{4\pi^2}{h^2}\right)eE\left(\frac{d^2E}{d\mathbf{k}^2}\right) \tag{3.210}$$

Comparing this equation with that for a free, classical particle where we have $m_e(dv/dt) = eE$ and $a = (dv/dt) = (eE/m_e)$, we then define the effective electron mass m_e^* as:

$$m_e^* = \left(\frac{h^2}{4\pi^2}\right)\left(\frac{d^2E}{d\mathbf{k}^2}\right)^{-1} \tag{3.211}$$

so that:

$$a = \left(\frac{eE}{m_e^*}\right) \tag{3.212}$$

From Equation 3.211, the effective mass is determined by $\left(\dfrac{d^2E}{d\mathbf{k}^2}\right)^{-1}$ (see Figure 3.105a). For a free electron $m_e^* = m_e$ because $E = \dfrac{\hbar^2\mathbf{k}^2}{2m_e}$ and $\dfrac{d^2E}{d\mathbf{k}^2} = \dfrac{\hbar^2}{m_e}$. All the equations for a free electron may be used for an electron in a crystal, provided that m_e in each case is replaced by the suitable m_e^*. For example, we may write:

$$E = \frac{\hbar^2\mathbf{k}^2}{2m_e^*} \tag{3.213}$$

Typically one considers states near the top of the valence band to be holes (particles of charge $+e$) with free electron-like dynamics but with effective mass m_h^*, and states near the bottom of the conduction band to be electrons with free electron-like dynamics but effective mass m_e^*. The effective mass is inversely proportional to the curvature of the band, and in general m^* is different in each direction of the crystal and is a tensor.

From experimental values for m_e^*, it is apparent that the effective mass need not always be larger than m_e. It can be smaller, and it may even be negative. For most metals, it is from one-half to twice m_e. For some transition metals, it is much higher than m_e, and for semiconductors it is lower. From Figure 3.105a, near $\mathbf{k} = 0$,

the effective mass approaches m_e. With **k** increasing, m_e^* also increases, and it reaches a maximum value at the inflection point ($\mathbf{k}_0$) of the E-**k** curve. Above the inflection point ($\mathbf{k} > \mathbf{k}_0$), m_e^* becomes negative, and as **k** approaches π/a it decreases to a small negative value. Near the bottom of the band, the effective mass m_e^* has a constant positive value because the quadratic equation $E \propto k^2$ holds over a small region here (second derivative is a constant). As **k** increases, the quadratic relation no longer holds, and m_e^* changes with **k**. Beyond $\mathbf{k}_0$, the mass m^* becomes negative because the region is close to the top of the band, and a negative mass is to be expected. The way to interpret this is to consider that for $\mathbf{k} > \mathbf{k}_0$ the velocity decreases, and therefore the acceleration is negative, implying a negative mass. In this region of **k**-space, the lattice exerts such a large retarding force on the electron that it overcomes the applied force and produces a negative acceleration. In other words, in the upper half of the band the electron behaves as a positively charged particle, referred to as a hole. Because we can describe the holes in terms of an E-**k** diagram, we can again define an effective mass simply by putting a minus sign in front of Equation 3.210. This turns out to be positive and given by the curvature of the E-**k** diagram at the top of the valence band. Therefore, here we finally have our answer why even metals might produce a positive Hall effect and why the conductivity is not simply proportional to the electron density; it matters where in the band these electrons find themselves!

Neither the electron nor the hole as described by its effective mass exists outside of the material; they are more properly referred to as quasiparticles. The fictitious positively charged particle is even stranger than a conduction band electron with an effective mass. If we add a hole to a completely filled valence band, we must end up with an empty electron state. Therefore, the hole must have a charge of $+e$ (compared with an electron's $-e$). If an empty state exists at low energy, then it is energetically favorable for an electron from a higher energy state to fall into it. In terms of holes, this means it is energetically favorable for the hole to rise to the top of the valence band, i.e., the energy scales for holes are reversed from those for electrons.

In Figure 3.105b we plot the degree of freedom of an electron, f_k, as a function of **k**. Here f_k, a measure of the extent to which an electron in state **k** is free, is defined as:

$$f_k = \frac{m}{m^*} = \frac{m}{\hbar^2}\left(\frac{d^2E}{d\mathbf{k}^2}\right) \tag{3.214}$$

If the effective mass is large, f_k is small, or the particle behaves as a heavy particle. When $f_k = 1$, the electron behaves as a free electron. In the lower half of the band, f_k is positive, and in the upper half f_k is negative.

The behavior of electrons and holes near the band edges determines most of the optical and electronic properties of a solid-state device. From the preceding, near the band edges, the electrons and the holes can be described by a simple effective mass picture, i.e., the electrons behave as if they are in free space except their masses are m_e^* and m_h^*, respectively. In Figure 3.106, we show a four-band model for a generic semiconductor with a parabolic approximation of the bands. This simplified band structure was first proposed by Kane (1957) and is valid near $\mathbf{k} = 0$. The valence band features a heavy holes (hh) band (I), a light holes (lh) band (II), and a split-off (so) band (III).

Electron transitions 1 and 2 are from the heavy holes band to the conduction band and from the light holes band to the conduction band, respectively. Split-off hole transitions are also possible.

As shown in Figure 3.106, along a given direction, the top two degenerate valence bands can be approximately fitted by two parabolic bands with different curvatures: the heavy holes band (the wider band,

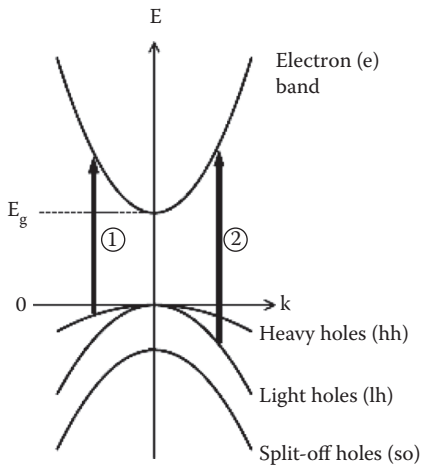


FIGURE 3.106 Four-band model. In the valence band, we have a heavy holes (hh) band (I), a light holes (lh) band (II), and a split-off (so) band (III).

with smaller $\frac{\partial^2 E(k)}{\partial k_i \partial k_j}$) and the light holes band (the narrower band, with larger $\frac{\partial^2 E(k)}{\partial k_i \partial k_j}$). Thus, the effective mass, in general, is tensorial with components:

$$\frac{1}{m^*_{ij}} \equiv \frac{1}{\hbar^2} \frac{\partial^2 E(k)}{\partial k_i \partial k_j} \tag{3.215}$$

which represents a generalization of Equation 3.211.

Example 3.5: In Si, $m_{hh} = 0.53\ m_0$ and $m_{lh} = 0.16\ m_0$; and in GaAs, $m_{hh} = 0.51\ m_0$ and $m_{lh} = 0.074\ m_0$.

Because there are multiple valence bands available for holes but there is only one conduction band for electrons, electromagnetic radiation absorption by holes is different from that of electrons. Free electrons located at wave vector $\mathbf{k} = 0$, at the bottom of a single parabolic conduction band, may reach higher energy states only if their momentum is increased. As a consequence, in n-type semiconductors, conduction band electrons may only be excited by the simultaneous absorption of a photon and the absorption or the emission of a phonon to conserve momentum (see also direct and indirect bandgap transitions discussed further below). This three-particle process (photon-electron-phonon) is less probable than a two-particle process (electron-photon). With p-type semiconductors, one can have such direct (i.e., no phonons needed) transitions between the different degenerate valence bands (heavy and light hole bands) because they occur at the same **k**. Because no momentum or **k** change is required, holes produce stronger free carrier absorption than conduction band electrons.

Below we take a closer look at the density of states (DOS) function, which describes the number of allowed energy states that are available in a system per unit energy and per unit volume (i.e., in units of number of states/eV/cm^3). From the preceding we can appreciate that the density of states, the effective mass, and the electron mobility are all correlated. When traveling through a crystal and bumped off course by a phonon or an impurity, the new energies an electron can adopt depend on the number of available states at the bottom of the conduction band. The density of states at the bottom of the conduction band is larger for Si than it is for GaAs, and as a consequence an electron traveling through silicon has a greater chance of being knocked off course into an allowed energy state. In general, then, a small effective mass is indicative of a relatively low number of energy levels at the bottom of the conduction band. The effective mass of an electron in silicon is six times heavier than that in GaAs, and as a consequence, based on Equation 3.9, the electron mobility in GaAs is six times larger or an electron can race six times faster through a GaAs lattice compared with a Si lattice (see Table 3.9). GaAs transistors, although much more difficult to fabricate than Si transistors, are used in cases where speed is of utmost importance; this includes military applications and the latest generations of supercomputers (e.g., the Cray 3).

Table 3.9 Electron and Hole Mobilities

Material	Mobility (cm^2/V-s)
Si	$\mu_n = 1500$, $\mu_p = 460$
Ge	$\mu_n = 3900$, $\mu_p = 1900$
GaAs	$\mu_n = 8000$, $\mu_p = 380$

Particle Distributions Functions

Introduction

Particle distributions functions *f(T, E)* represent the probability that a particular state with energy *E* is occupied by a particle (say an electron) in equilibrium at a given temperature *T*. Earlier in this chapter we introduced the Maxwell-Boltzmann distribution, a classical distribution of particles, illustrated in Figure 3.6. In this distribution function, it is assumed that all the particles are distinguishable. This kind of consideration comes from the fact that all particles have characteristic wave properties according to the de Broglie hypothesis. Two particles can be considered to be distinguishable if their separation *d* is large compared with their de Broglie wavelength and are considered to be indistinguishable if their wave packets overlap significantly. In such case, a Fermi-Dirac particle distribution and (f_{FD}) must be applied. The thermal de Broglie wavelength is roughly the average de Broglie wavelength

of the gas particles in an ideal gas at the specified temperature and is given by:

$$\lambda_{DB} = \sqrt{\frac{2\pi\hbar^2}{mk_B T}} \sim \frac{h}{p} \tag{3.216}$$

We can take the average interparticle spacing in an ideal gas to be approximately $(n = V/N)^{1/3}$, where V is the volume, N is the number of particles, and n is the density of particles. When the thermal de Broglie wavelength is much smaller than the interparticle distance, the gas can be considered to be a classical or a Maxwell-Boltzmann gas, or:

$$\lambda_{DB} <<< d = n^{-\frac{1}{3}} \tag{3.217}$$

On the other hand, when the thermal de Broglie wavelength is on the order of, or larger than, the interparticle distance, quantum effects will dominate, and the gas must be treated as a Fermi gas or a Bose gas, depending on the nature of the gas particles. Particles become indistinguishable when $d = n^{-1/3} \sim \lambda_{DB}$. The de Broglie temperature is given by:

$$T_{DB} = \frac{2\pi\hbar^2}{mk_B} n^{\frac{2}{3}} \tag{3.218}$$

Example 3.6: For an electron gas in metals, $n = 10^{22}$ cm^{-3}, $m = m_e$, and $T_{DB} \sim 3 \times 10^4$ K, but for a gas of Rb atoms, $n = 10^{15}$ cm^{-3} and $m_{atom} = 10^5 m_e$, so we obtain a $T_{DB} \sim 5 \times 10^{-6}$ K. In other words, at room temperature electrons are indistinguishable with overlapping wave functions, whereas Rb atoms only become so at very low temperatures.

Besides the Fermi-Dirac distribution, we will also introduce the Bose-Einstein distribution function (f_{BE}) and explain under what conditions they each apply. We will see that at $T < T_{DB}$, f_{BE} and f_{FD} are strongly different from f_{MB} and at $T >> T_{DB}$: $f_{BE} \approx f_{FD} \approx f_{MB}$.

Fermi-Dirac

When assigning electrons to the energy levels one must require, as Sommerfeld did, that the allowed wave functions obey Pauli's principle (introduced in 1925), i.e., one can only put two electrons with opposite spin in each level of quantum number n. When that is done this level is filled. One then proceeds to the next higher level for the next pair of electrons. The result is obvious—all the lowest levels are filled with pairs of electrons until they reach some maximum value of energy E_F, the so-called Fermi energy. The exclusion principle applies to all "spin one-half" particles, which include electrons, protons, and neutrons. If we draw the distribution function, i.e., the probability of filling a level, $f(E,T)$, as a function of E, for two different temperatures, we find the result shown in Figure 3.107.

The probability of an energy state being occupied is called the Fermi factor, $f(E,T)$. As T approaches zero, the Fermi-Dirac distribution becomes a step function. At 0 K, f(E, T) = 1 until we reach the maximum level E_F, after which it falls to zero. If we increase the temperature, thermal energy can excite electrons to energy levels higher than E_F. Because the kinetic energy of the lattice ions is of the order of kT (~0.025 eV), electrons cannot gain much more than kT in collisions with lattice ions. This is determined by the Fermi-Dirac distribution, which applies for any particle that follows Pauli's exclusion principle—no more than two particles of opposite spin being allowed in a given energy level. Half-integer spin particles are called fermions, and Fermi-Dirac statistics and Pauli's exclusion principle hold. The "material" particles, such as electrons, protons, and neutrons, are all fermions, and without Pauli's exclusion principle, the plethora of chemical elements and the variety of our physical world would simply not exist. Other particles such as α-particles, deuterons, photons, and mesons do not obey Pauli's exclusion principle. Such particles are called bosons

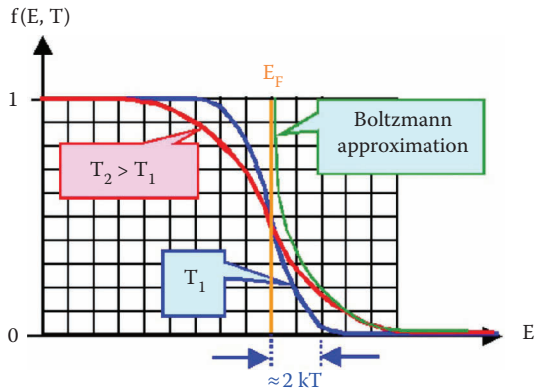


FIGURE 3.107 The Fermi-Dirac distribution function for two different temperatures. The Boltzmann approximation is indicated as well (green line).

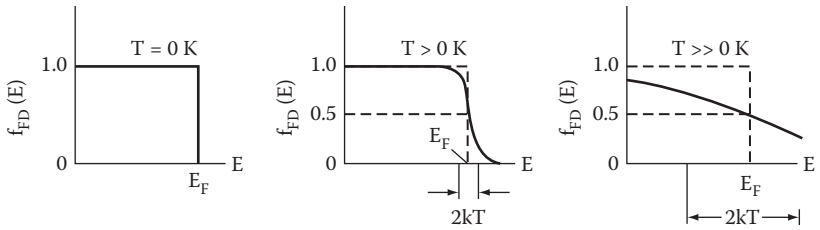


FIGURE 3.108 Evolution of Fermi distribution function as a function of temperature. At the highest temperatures the Fermi-Dirac function smears out and starts resembling a Boltzmann distribution.

and have either zero intrinsic spin or integral spin quantum numbers. Fermion and boson particle distributions are compared further below in Figure 3.108. The wave function that describes a collection of fermions must be antisymmetric with respect to the exchange of identical particles. One fermion of a system in a certain state prevents all other fermions from being in that state. The mathematical expression for the Fermi-Dirac distribution as a function of energy and temperature is:

$$f(E, T)_{FD} = \frac{1}{e^{\left[\frac{(E-\mu)}{kT}\right]}+1} \text{ or } \frac{1}{Ae^{\frac{E}{kT}}+1} \qquad (3.219)$$

where μ is the chemical potential and $A = e^{-\mu}$. At T = 0, one can see that $\mu(T = 0) = E_F$, the Fermi energy. Above T = 0 there is no abrupt energy that separates filled from unfilled levels, so the definition of the Fermi energy must be slightly modified. Note that when f(E, T) = 1/2, there is a 50-50 probability of finding the level occupied and $E = E_F$. Thus, at temperature *T*, the Fermi energy is defined as that energy for which the probability of being occupied is 0.5. The Fermi energy plays a very important role in the band theory of semiconductors and metals. In metals the Fermi energy is an effective cutoff level for the allowed energies of the electrons. By analogy, imagine a sea of electrons with a "depth" of E_F. At room temperature only a small fraction of the electrons will ever have an energy much above that sea level ($\approx 2kT_1$ at T_1). The energy E_F corresponds to the chemical potential, μ, i.e., the amount of energy needed to add an electron to the system ($E_F = \mu$). For fermions, the chemical potential may be either positive or negative. From thermodynamics, the chemical potential, and thus the Fermi energy, is related to the Helmholtz free energy: $\mu = F(n + 1) - F(n)|_{T,V}$, where F = U − TS. For comparison, we have indicated the Maxwell-Boltzmann distribution in Figure 3.107 as well (green line).

With $T > 0$ K, the $f(E, T)$ function has the general shape as sketched in Figure 3.108, but as the temperature keeps on increasing it gradually smears out, and finally at very high temperature ($T \gg 0$ K) it begins to look like an ordinary Boltzmann distribution as shown in Figure 3.108 (see also Figure 3.6). At ordinary temperatures, say 1000 K, kT = 0.088 eV, whereas for gold E_F = 5.51 eV and for sodium it is 3.12 eV-typical values for metals. In other words, $E_F \gg kT$, and an electron gas in a metal cannot usually be treated as an ordinary Maxwell-Boltzmann gas. With $E_F \gg kT$, the electron distribution is quantum-mechanical, and the electron gas is said to be degenerate. A Fermi gas well described by a Fermi-Dirac distribution that is approximately step-like is termed degenerate.

Remember that Drude used a Maxwell-Boltzmann distribution for the free electrons in a metal. From the above we now recognize that this was the wrong particle distribution to use. The Maxwell-Boltzmann distribution, shown in Figure 3.6, is a classical distribution of particles, but with particles satisfying quantum statistics, a Fermi distribution must be used. The electronic velocity vector component distribution for a single direction v_i according to Drude (based on Equation 3.17, which gives the distribution of speeds of molecules rather than their component velocities), is given as:

$$f_{MB}(v_i) = n\left(\frac{m}{2\pi k_B T}\right)^{\frac{3}{2}} e^{-\frac{\frac{1}{2}mv_i^2}{kT}} \qquad (3.220)$$

where n = N/V with *n* the electron density, *V* the volume, and *N* the sum of all electrons. The Maxwell-Boltzmann distribution assumes that the described particles are distinguishable. In other words, we can make a distinction as to which particle is in which energy state. For an ideal gas, that is certainly the case,

but this is not true for electrons, and this is the reason why Drude's model ultimately failed. Sommerfeld used the same classical gas as Drude but used the quantum Fermi-Dirac distribution for the electrons' velocity vector component in one direction (v_i):

$$f_{FD}(v_i) = \frac{(m/\hbar)^3}{4\pi^3} \frac{1}{\exp\left[\left(\frac{1}{2}mv_i^2 - kT_0\right)/kT\right] + 1} \tag{3.221}$$

The temperature T_0 is determined by the normalization $n = \int f(v_i)dv_i$.

We can understand now why the electron gas does not contribute to the heat capacity of metals. As a metal is heated, the only way an electron can take up energy is by moving into a somewhat higher allowed energy level. But a typical electron is buried deep inside the Fermi sea, and there are no empty levels to move to because each level above the electron is already occupied by a pair of electrons of opposing spin. Only the relatively few electrons at the top of the distribution can find empty levels to move into. These electrons find themselves in the so-called Maxwellian tail of the Fermi-Dirac function. They are the only electrons that contribute to the heat capacity, as we will calculate below. Thus, at ordinary temperatures the electronic heat capacity is almost negligible, as we can see from Figure 3.16. But why then can all the electrons in a free electron gas contribute to the electrical conductivity? We would expect that also in this case, electrons take energy from the electrical field and move into higher energy levels. What happens in this case, however, is a shift of the entire set of electron energy levels from lower to higher values. The Fermi distribution, in this case, as we will show below in Figure 3.121, is shifted bodily by the applied field. Thus, electrons can acquire an average drift velocity in the field without violating Pauli's exclusion principle.

Bose-Einstein Distribution

Bosons are a third type of particles that come with their own kind of distribution law, i.e., the Bose-Einstein distribution. Violations of the Maxwell-Boltzmann statistics are observed if the density of particles is very large (neutron stars), the particles are very light (electrons in metals, photons), or they are at very low temperatures (liquid helium). Classical and quantum mechanics particle distributions are compared in Table 3.10. From this table we recognize that with a very large E or small T all three distributions reduce to the classical Maxwell-Boltzmann form.

Whole-integer spin particles are called bosons, for which Bose-Einstein statistics are applicable. The wave function that describes a collection of bosons must be symmetric with respect to the exchange of identical particles. One boson of a system in a certain state increases the probability of finding another boson in this state. Satyendra Bose and Albert Einstein developed these statistics in the

TABLE 3.10 Classical and Quantum Distributions of Particles, with $\mu = E_F$ the Chemical Potential of the Particle

Distribution Name	Properties of the Particles	Examples	Distribution Function, F
Maxwell-Boltzmann	Spin does not matter. Unlimited number of particles per state. Particles are identical but distinguishable. Wave functions do not overlap.	Classical gas. Fermions and bosons at high T ($\mu - E >> kT$).	$F_{MB} = e^{\left(\frac{\mu - E}{kT}\right)}$ or $Ae^{\left(\frac{-E}{kT}\right)}$
Bose-Einstein	Boson particles are indistinguishable with integer spin (0, 1, 2…). Unlimited number of particles per state. Wave functions overlap.	Liquid ^{4}He, photons, Cooper pairs, excitons, Rb.	$F_{BE} = \frac{1}{e^{\left(\frac{E - \mu}{kT}\right)} - 1}$ or $\frac{1}{Ae^{\frac{E}{kT}} + 1}$
Fermi-Dirac	Fermion particles are identical with half-integer spins (1/2, 3/2, 5/2…). Wave functions overlap. Never more than one particle per state.	Free electrons in metals, protons, and neutrons; electrons in white dwarfs.	$F_{FD} = \frac{1}{e^{\left(\frac{E - \mu}{kT}\right)} + 1}$ or $\frac{1}{Ae^{\frac{E}{kT}} + 1}$

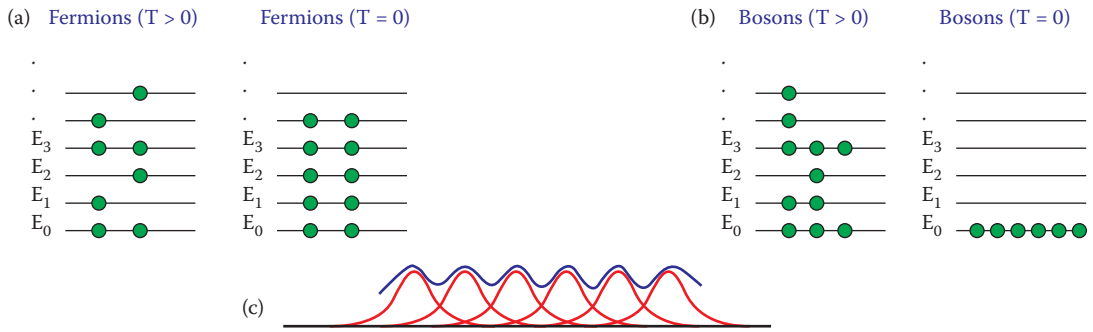


FIGURE 3.109 (a) Fermions at $T > 0$ and at T = 0 filling up a set of energy levels. (b) Bosons at $T > 0$ and T = 0. The lowest state is macroscopically populated. (c) The particles get so close together that their wave functions overlap, and the condensate is described by a single, macroscopic ψ.

1924–1925 time frame. Bose, a reader in physics at Dacca University, first derived Planck's blackbody law anew from a combination of light quanta of zero mass and statistics. By treating radiation as a quantum gas and counting particles instead of wave frequencies, Bose cut quantum theory loose from its classical antecedents. Then Einstein developed these statistics further and applied them to collections of atoms—gas or liquid—obeying the same rules, and he predicted the phenomenon of what later would be called the Bose-Einstein condensation (BEC) or *superatoms*. If one concentrates a large number of identical bosons in a small region at low temperatures, their wave functions start to overlap, and the bosons lose their individual identities and become one object. The BEC is a macroscopic matter wave, and it is formed of atoms that are delocalized and indistinguishable. In other words, order appears in momentum space (atoms adopt a collective behavior). In Figure 3.109a and b we compare how fermions and bosons fill up a set of energy levels and in c we illustrate how bosons, when they get close together, can be described by a single macroscopic wave function ψ. Einstein was himself not convinced that BEC could happen: "*Die Theorie ist schön, aber ist sie auch war*?" The Bose-Einstein distribution is given as:

$$f_{BE}(E) = \frac{1}{e^{\left(\frac{(E-\mu)}{kT}\right)} - 1} \text{ or } \frac{1}{Ae^{\frac{E}{kT}} - 1} \qquad (3.222)$$

where A = exp(−μ). The boson statistics would mark Einstein's last great contribution to quantum theory.

Bose-Einstein, Maxwell-Boltzmann, and Fermi-Dirac distributions are compared in Figure 3.110 for the same value of A = 1. It is clear that these

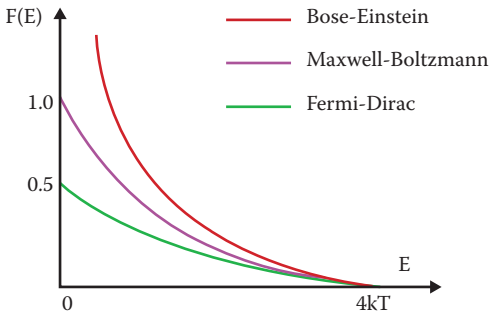


FIGURE 3.110 Bose-Einstein, Maxwell-Boltzmann, and Fermi-Dirac distributions compared.

distributions deviate the most at low temperatures, so one expects to see quantum phenomena emerge at low temperatures.

With bosons there is no exclusion principle, and a macroscopic number of bosons occupy the lowest energy quantum state. The blackbody law, for example, is a direct result of photons, which are bosons, all trying to get into the same energy state. Among other things, bosons also explain superconductivity, lasers (detailed in Chapter 5), and superfluidity in ^{4}He. A superfluid has no viscosity; in effect it experiences no friction and is able to flow forever while at the same time the thermal conductivity becomes very large!

Fermi Surfaces, Brillouin Zones, Density of State Functions, and Conductivity as a Function of Quantum Confinement

Introduction

In this section we consider how electronic processes, like charge transport, are affected by the dimensions of the solid. For this, we need to introduce the density of state functions for electronic systems with different levels of quantum confinement. The

density of state function, $G(E)$, is also called the *degeneracy* of a system and is often abbreviated as DOS. The DOS function describes the number of allowed energy states that are available in a system per unit energy and per unit volume (i.e., in units of number of states/eV/cm^3). Earlier, we showed that for a gas of molecules, the number of molecules per unit volume, i.e., density $n(E)$, that have energies between E and $E + dE$, is $n(E)dE$, where $n(E) = G(E)f_{MB}(E)$ with $f_{MB}(E)$, the Maxwell-Boltzmann distribution, representing the probability function of occupancy of a state with energy E. In the case of an electron gas, the number of electrons, $n(E)$, with energies between E and $E + dE$, is again given by the product of $G(E)f_{FD}(E)$ and a distribution function $f_{FD}(E)$, which in this case is the Fermi-Dirac distribution. Thus, combining the density of available states (DOS) with the Fermi function, one can calculate the density of filled states $n(E)$. In other words, we obtain a number for those charge carriers that can contribute to conductivity and charge transport.

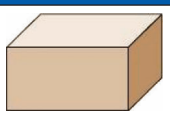
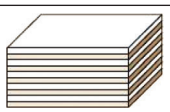
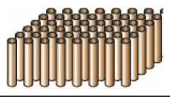
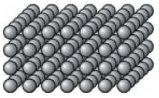
Systems with 3, 2, 1, and 0 degrees of freedom for electrons to move in are referred to as bulk material, quantum wells, quantum wires, and quantum dots, respectively, as summarized in Table 3.11. In this table, λ_{DB} represents the de Broglie wavelength (see Equation 3.216). It is the size of L_1, L_2, L_3 compared with λ_{DB} that determines whether we are dealing with a bulk semiconductor (L_1, L_2, $L_3 > \lambda_{DB}$), a quantum well (L_1, $L_2 > \lambda_{DB} > L_3$), a quantum wire ($L_1 > \lambda_{DB} > L_2$, L_3), or a quantum dot ($\lambda_{DB} > L_1$, L_2, L_3). We will see that the DOS function, $G(E)$, strongly depends on the degrees of freedom of the electrons in the system.

To calculate the charge carrier transport or current density ($\mathbf{J} = -ne\mathbf{v}$), using solids of different dimensionalities, we need to know the charge carrier density n as a function of dimensionality, but also the concept of Fermi surfaces needs to be introduced, and we need to get reacquainted with the Brillouin zones of Chapter 2. A Fermi surface is a surface of constant energy E_F in **k**-space, and at absolute zero temperature, it separates occupied from unoccupied quantum states. The way Fermi surfaces fill Brillouin zones determines whether a material will be a metal, semimetal, semiconductor, or an insulator. For example, if the first Brillouin zone is completely filled and there is a large gap between it and the next Brillouin zone, we have an insulator; a smaller gap makes for a semiconductor. If the first band is not completely filled or overlaps with an empty second Brillouin zone, a conductor or a semimetal results, respectively.

Bulk Materials: L_1, L_2, $L_3 > \lambda_{DB}$

Fermi Surface for Bulk Materials A Fermi surface is a surface of constant energy E_F in **k**-space, and at absolute zero temperature, it separates occupied from unoccupied quantum states. With a large number of electrons N at T = 0, the electrons will occupy the lowest energy states, consistent with the exclusion

Table 3.11 Examples of Reduced-Dimensional Material Geometries and Definitions of Their Dimensionality and of the Associated Type of Confinement*

$L_{1,2,3} > \lambda_{DB}$	No nanostructures	No confinement	Bulk material	
$L_{1,2} > \lambda_{DB} > L_3$	2D nanostructures	1D confinement	Wells	
$L_1 > \lambda_{DB} > L_{2,3}$	1D nanonstructures	2D confinement	Wires	
$\lambda_{DB} > L_{1,2,3}$	0D nanostructures	3D confinement	Dots	

*Here λ_{DB} represents the de Broglie wavelength.

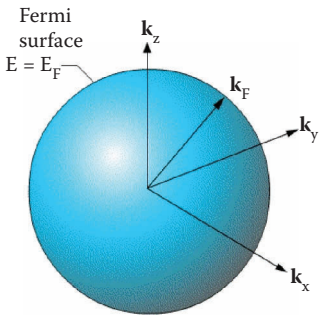


FIGURE 3.111 Free electron Fermi surface.

principle. Thus, the N electrons will fill up the lowest $N/2$ energy levels (two electrons per level). The energy of the last filled (or half-filled) level at T = 0 is called the Fermi energy, E_F. An unrestricted 3D Fermi electron gas is isotropic, so that a surface of constant energy E in **k**-space is a sphere with a surface called the Fermi surface (Figure 3.111). This surface separates the occupied from the unoccupied states at 0 K. At T = 0 K, all the free electron states are occupied up to an energy E_F, and all fall within a Fermi sphere with a Fermi wave vector, $\mathbf{k}_F$. When electrons are not free, a Fermi surface still exists, but it is distorted by the interaction with lattice ions (see, for example, Figure 3.117). The Fermi energy E_F is an effective cutoff level for the allowed energies of the electrons and corresponds to the chemical potential, μ, i.e., the amount of energy needed to add an electron to the system ($E_F = \mu$). The quantities $\mathbf{k}_F$ and E_F are called the Fermi wave vector and Fermi energy, respectively, in honor of Enrico Fermi. To calculate their exact value, we need to introduce the density of states function (see below, Equations 3.238 and 3.232, respectively).

Brillouin Zones for Bulk Materials Brillouin zones for crystals were first introduced in Chapter 2 in the context of x-ray diffraction. An x-ray diffraction pattern of a crystal represents a map of the reciprocal lattice, a Fourier transform of the lattice in real space, or also a representation of the lattice in **k**-space. Each Brillouin zone is a primitive cell of the reciprocal lattice, which corresponds to the Wigner-Seitz primitive cell of the real lattice (Figure 2.34). The boundaries of Brillouin zones satisfy the von Laue conditions for diffraction, and the Brillouin zone surface describes all **k** vectors that are constructively diffracted by the crystal. In other words, Bragg planes bound the Brillouin zones. This is true for the x-rays considered in Chapter 2 but also for the matter waves associated with electrons, neutrons, etc., discussed in this chapter. In discussing the Ewald sphere in Chapter 2 we also pointed out that the Bragg planes bisect the shortest vectors of the reciprocal lattice and that one can define the first Brillouin zone as the set of points in **k**-space that can be reached from the origin without crossing any Bragg planes. The second Brillouin zone is the set of points that can be reached from the first zone by crossing only one Bragg plane. The nth Brillouin zone can be defined as the set of points that can be reached from the origin by crossing $n-1$ Bragg planes, but no fewer (see the Brillouin zone for a square and a triangular 2D lattice in Figure 2.33). To get reacquainted with the procedure of constructing Brillouin zones, we start off here with a simple square lattice of atoms with interatomic distance, a. Its reciprocal lattice is also a square, with reciprocal lattice base vector of length $2\pi/a$. That reciprocal lattice is shown in Figure 3.112a, where we also have drawn the three shortest vectors marked $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$. Three lines, Bragg planes, are drawn as perpendicular bisectors of these three vectors. By drawing all the lines equivalent by symmetry with those bisector lines, we obtain the regions in **k**-space that constitute the first three Brillouin zones as marked (Figure 3.112b).

We turn now to a 3D crystal where we find again that a zone structure exists that may be different for different directions in **k**-space. The coordinates $\mathbf{k}_x$, $\mathbf{k}_y$, and $\mathbf{k}_z$ specify a point in **k**-space: a value of the vector from the origin that specifies the momentum of the electron. The zones in **k**-space corresponding to allowed energies for motion of an electron in a solid are 3D Brillouin zones forming nested sets of polyhedra. Different potentials exist in different directions, so the electron wavelength and crystal momentum, $\mathbf{k} = 2\pi/\lambda$, differ with direction, and many different parabolic E-**k** relationships exist, depending on the crystalline momentum. For a simple example, consider a cubic crystal with a Bragg reflection (Equation 2.20 with d = a) occurring when:

$$\mathbf{k}_x = \frac{2\pi n_x}{a},\ \mathbf{k}_y = \frac{2\pi n_y}{a},\mathbf{k}_z = \frac{2\pi n_z}{a},$$
$$n_x, n_y, n_z, = \pm 1, \pm 2, \pm 3, \ldots \qquad (3.223)$$

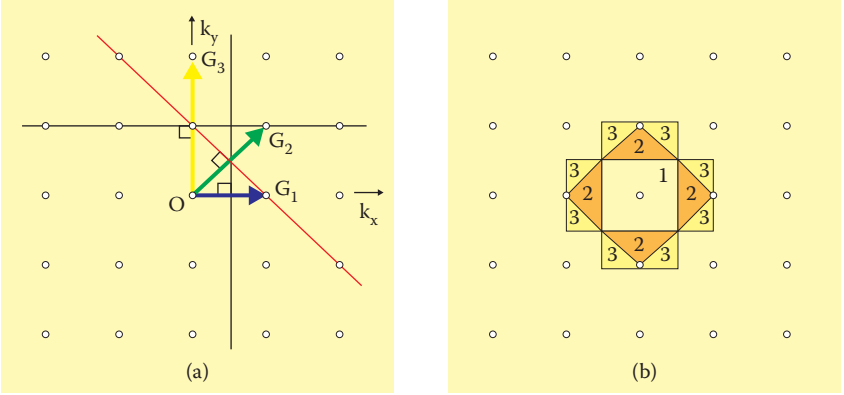


FIGURE 3.112 Illustration of the definition of the Brillouin zones for a 2D square Bravais lattice. To go from (a) to (b): the numbers in (b) denote the zone to which the region belongs. The numbers are ordered according to the length of the vector **G** used in the construction of the outer boundary of the zone (a).

These are the zone boundary values corresponding to discontinuities in the allowed energy levels. From Bragg's law, reflecting planes with the largest interplanar spacings (d = a) have the smallest values for **k**. The planes with the largest spacing are the {100} planes. Thus, the first Brillouin zone in **k**-space is bounded by the {100}* planes at $\mathbf{k}_x = \frac{\pm\pi}{a}$, $\mathbf{k}_y = \frac{\pm\pi}{a}$, $\mathbf{k}_z = \frac{\pm\pi}{a}$, which therefore is a cube with faces π/a from the origin as shown in Figure 3.113. The next set of reflecting planes are the {110} planes with spacing $d = a/2^{1/2}$. Corresponding planes in **k**-space are:

$$\pm\mathbf{k}_x \pm \mathbf{k}_y = \frac{2\pi}{a}, \quad \pm\mathbf{k}_x \pm \mathbf{k}_z = \frac{2\pi}{a}, \quad \pm\mathbf{k}_x \pm \mathbf{k}_z = \frac{2\pi}{a} \tag{3.224}$$

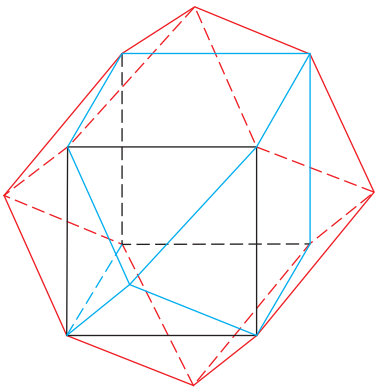


FIGURE 3.113 First two Brillouin zones for a simple cubic lattice. The first zone is the cubic volume $8\pi^3/a^3$ in **k**-space. The second zone is the **k**-space between the cube and the circumscribed dodecahedron.

* See Chapter 2 on crystallography on Miller indices.

These 12 planes outline the dodecahedron shown in the same figure. The second Brillouin zone is the **k**-space between the cube and the dodecahedron.

In Figure 3.114 we show what happens when one fills up the first Brillouin zone (BZ) of a face-centered cubic structure with electrons. In this case the largest spacing that satisfies the Bragg condition involves both {111} and {200} planes. These are the set of planes that would be reached first (largest a is smallest **k**). The zone boundaries then derive partly from planes derived from {200} and partly from planes derived from {111}. Envision a growing Fermi sphere inside this first Brillouin zone. Near the center of the zone the electrons are virtually free and behave like an electron gas with an energy given by Equation 3.144. At these energies, well below the zone boundaries, the surfaces of constant energies are represented as spheres within the Brillouin zone as drawn in Figure 3.114a. But when adding more and more electrons, we approach the diffraction

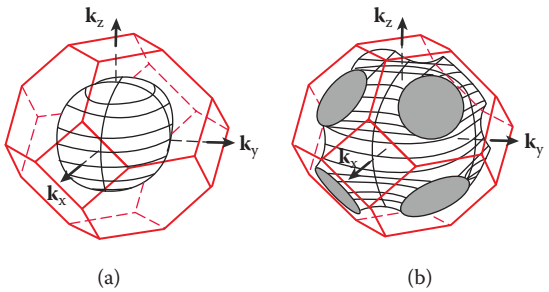


FIGURE 3.114 (a) The first Brillouin zone of a face-centered cubic (FCC) structure with surfaces of constant energy of electrons shown for nearly free electrons near the bottom of the zone and (b) electrons at the zone boundary.

boundaries, and the wave vectors near or at the BZ feel the periodic potential of the crystal, whereas the others do not. All wave vectors that end on a BZ will fulfill the Bragg condition and thus are diffracted. Because electrons cannot transgress this boundary, the spherical surface becomes distorted by bulging out toward the zonal plane (Figure 3.114b). These electrons no longer behave as a free electron gas. Wave vectors completely in the interior of the first BZ, or between any two BZs, will never get diffracted; they move pretty much as if the potential would be constant; i.e., they behave very close to the solutions of the free electron gas.

If the first Brillouin zone is completely filled and there is a considerable gap between it and the next Brillouin zone, we have an insulator at hand. With a smaller gap, we are dealing with a semiconductor. If the first band is not completely filled or overlaps with an empty second Brillouin zone, we are dealing with a conductor or a semimetal, respectively.

Some other jargon often associated with the difference between a metal, a semiconductor, and an insulator is the characterization of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). If the difference in the energy between the HOMO and the LUMO is zero, then just a little energy can promote an electron in an unoccupied level. Therefore, with an electrical potential difference, some electrons are very mobile and give rise to electrical conductivity in a metal. When the temperature increases, there are more electrons excited toward empty orbitals. However, the conductivity decreases because the vibration of the nuclei increases the collisions between the transported electrons and the nuclei, so there is a less-efficient transport or the resistance increases. When the HOMO-LUMO energy difference is nonzero, there is an electronic gap. If the bandgap is small, thermal excitations can promote electrons to unoccupied levels; consequently, those electrons can contribute to the electrical conductivity. This is the case of semiconductors, and that is why the conductivity of semiconductor increases with temperature. Insulators are characterized by a huge HOMO-LUMO energy difference, and the electrons cannot reach the unoccupied levels; there is no measurable conductivity.

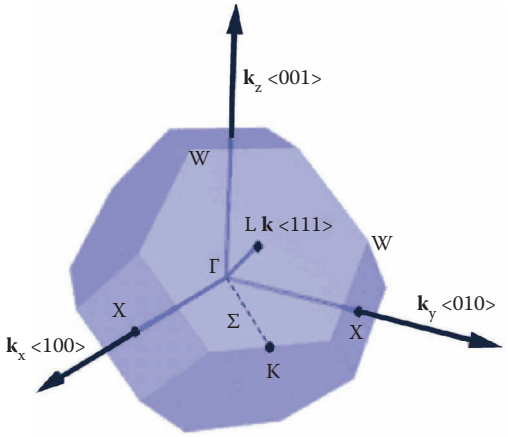


FIGURE 3.115 First Brillouin zone for Si. Points of high symmetry on the Brillouin zone have specific importance. The most important point for optoelectronic devices is the center at $\mathbf{k} = 0$, known as the gamma point Γ (000). One finds *X* at the surface boundaries (100), (001), and (010); *K* at (110); and *L* at (111). *K* is in the middle of an edge joining two hexagonal faces, *L* is at the center of a hexagonal face, and *W* is a corner point. Σ means directed from Γ to *K*.

In Figure 3.115 we show the first Brillouin zone for Si, a semiconductor. Letters are used to mark many of the high symmetry points on this first BZ boundary. The gamma point (Γ) is always the zone center, where $\mathbf{k} = 0$ ($\mathbf{k}_x = \mathbf{k}_y = \mathbf{k}_z = 0$), the *X* point is at $\mathbf{k}_x = 2\pi/a$ and $\mathbf{k}_y$ and $\mathbf{k}_z = 0$ (center of a square face), and the *L* point at $\mathbf{k}_x = \mathbf{k}_y = \mathbf{k}_z = \pi/a$ (center of a hexagonal face) (*a* is the lattice constant, i.e., cube edge). *K* is in the middle of an edge joining two hexagonal faces, *L* is at the center of a hexagonal face, and *W* is a corner point. Σ means directed from Γ to K. Most semiconductors have band edges of allowed bands at one of these points. Zone edges or surfaces are marked with symbols from the Roman alphabet, whereas the interior is marked with symbols from the Greek alphabet. Two example Greek letters not yet introduced are: Δ which means directed from Γ to *X* and Λ which means directed from Γ to L. These symbols are not marked in Figure 3.115 where we only introduced Γ and Σ but they are used in Figure 3.116. In Figure 3.116 the energy bands in 2D are plotted along the major symmetry directions in the first BZ for two of the industrially most important semiconductors, namely, Si and GaAs.

Group IV and III–V semiconductors have a band structure that appears somewhat similar because their basic character is controlled by sp^3 hybridization and tetrahedral bonding. From Figure 3.116a, in the

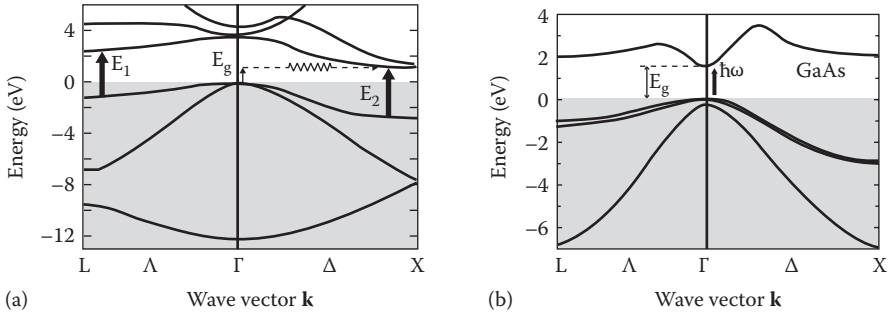


FIGURE 3.116 (a) Si is an indirect semiconductor because the maximum of the valence band (E_v at Γ) does not coincide with the minimum of the conduction band (E_c at X). For silicon, the valence band maximum (VBM) occurs at $\mathbf{k} = 0$ (Γ point). However, the conduction band minimum (CBM) occurs to the left of *X*. This makes Si an indirect bandgap material. The shortest distance between E_c and E_v is the bandgap E_g of Si: 1.1 eV (a). GaAs is a direct bandgap semiconductor because the VBM and CBM occur at the same point in the Brillouin zone, and it features a strong absorption. For GaAs, $E_g = 1.5$ eV (b).

case of silicon, the valence band maximum (VBM) occurs at $\mathbf{k} = 0$ (Γ point). However, the conduction band minimum (CBM) occurs to the left of *X*. Thus, the VBM and the CBM occur at different points in the Brillouin zone, and this makes Si an indirect bandgap material (with a bandgap $E_g = 1.1$ eV). For an electron to move from the VBM to the CBM, its **k** vector needs to be changed, making the process less likely because a phonon is required for the electron to change its momentum (three-particle process: electron, photon, and phonon). Its indirect bandgap makes Si a weak absorber of light and thus a poor optoelectronic material. The Si band diagram also has two critical points at E_1 (3.2 eV) and at E_2 (4.3 eV). Critical points (van Hove singularities) occur whenever $dE/dk = 0$, in other words, when the bands are parallel to each other. At those critical points, band transitions are more likely because no momentum change is required (parallel band effect). The Si band structure also features six equivalent conduction band minima at *X* along six equivalent <100> directions, and the valence band maxima for the heavy-hole, light-hole, and split-off bands are located exactly at the Γ point. Germanium and diamond also are indirect bandgaps. GaAs is a direct bandgap semiconductor (Figure 3.116b) because the VBM and CBM occur at the same point in the BZ, and it features a strong absorption at $\hbar\omega > E_g$ (for GaAs, $E_g = 1.5$ eV), making it an excellent optoelectronic material.

The Fermi surface for Cu, shown in Figure 3.117, is a sphere entirely contained within the first Brillouin zone. Even if the free electron Fermi sphere does not intersect a BZ boundary, its shape can still be affected at points close to the boundary where the energy bands begin to deviate from the free electron parabolic shape. This is the case with Cu, where in the <111> directions, contact is made with the hexagonal Brillouin zone faces. Eight short "necks" reach out to touch the eight hexagonal zone faces. The Fermi surface shown in Figure 3.117 can extend throughout many unit cells if the necks in the <111> directions are joined together with similar surfaces in adjacent cells. A section of such a continuous zone structure is shown in Figure 3.118. A magnetic field may close off these necks, a phenomenon that may be studied using cyclotron resonance. These necks are clearly evident from de Haas-van Alphen oscillations (http://www.lanl.gov/orgs/mpa/nhmfl/users/pages/deHaas.htm and http://physics.binghamton.edu/Sei_Suzuki/pdffiles/Note_dHvA.pdf) for magnetic fields in the <111> directions, which contain two periods, determined by the extremal "belly" (maximum) and "neck" (minimum) orbits. The de Haas-van Alphen

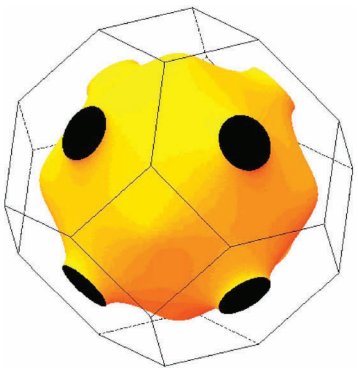


FIGURE 3.117 Brillouin zone for Cu with free electron sphere bulging out in the <111> directions to make contact with the hexagonal zone faces.

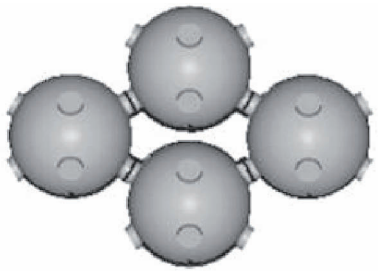


FIGURE 3.118 Continuous Fermi surfaces extending through adjacent unit cells in gold structures.

effect has its origin in the Bohr-Sommerfeld quantization of the orbits of conduction electrons under the influence of a magnetic field. A measurement of the temperature dependence of the oscillation amplitude permits a determination of the cyclotron frequency or, equivalently, the electron mass.

Density of States for Bulk Materials The DOS function $G(E)$ is a property that quantifies how closely packed energy levels are in some physical system. It is usually expressed as a function of internal energy E, $[G(E)]$, or as a function of the wave vector $\mathbf{k}$, $[G(\mathbf{k})]$. The density of $\mathbf{k}$-states, $G(\mathbf{k})$, in the Fermi sphere drawn in Figure 3.111 is the number of allowed states between $\mathbf{k}$ and $\mathbf{k} + d\mathbf{k}$, i.e., the number of states between a sphere of radius $\mathbf{k}$ and a sphere of radius $\mathbf{k} + d\mathbf{k}$ (analogous to the calculation of the density of states for gas molecules illustrated in Figure 3.7). In three dimensions the volume between the two shells, $\mathbf{V}$ is given by:

$$\mathbf{V}d\mathbf{k} = 4\pi\mathbf{k}^2 d\mathbf{k} \tag{3.225}$$

The density of states $G(\mathbf{k})_{3D}$ is then derived by dividing this volume by the volume of a single energy state. From Figure 3.66b we recognize that each value of $\mathbf{k}$ occupies a volume $V = (2\pi/L)^3$, so that the number of states per unit volume of $\mathbf{k}$-space is $1/V$ or $(L/2\pi)^3$. We also introduce a factor of two here from Pauli's exclusion principle: each energy state can accommodate two electrons, or:

$$G(\mathbf{k})_{3D}\,d\mathbf{k} = 2\frac{4\pi\mathbf{k}^2 d\mathbf{k}}{\left(\frac{2\pi}{L}\right)^3} = \frac{L^3\mathbf{k}^2}{\pi^2}d\mathbf{k} \tag{3.226}$$

To obtain the density of states in terms of the energy $[G(E)_{3D}]$, we use the relation between E and $\mathbf{k}$, derived earlier as $\mathbf{k} = \left(\frac{2m_e^* E}{\hbar^2}\right)^{1/2}$ (Equation 3.150). For every $\mathbf{k}$-state there is a corresponding energy state. Differentiating the latter expression with respect to energy leads to:

$$d\mathbf{k} = \left(\frac{2m_e^* E}{\hbar^2}\right)^{-\frac{1}{2}} \frac{m_e^*}{\hbar^2} dE \tag{3.227}$$

The density of energy states $G(E)_{3D}$, the number of allowed states between E and $E + dE$, is then obtained using the chain rule and dividing by L^3 (=V) because we want an expression per unit energy and unit volume:

$$G(E)_{3D} = G(\mathbf{k})_{3D}\frac{d\mathbf{k}}{dE} = \frac{k^2}{\pi^2}\frac{d\mathbf{k}}{dE} \tag{3.228}$$

Thus, we calculate the density of states for a parabolic band in a bulk material (three degrees of freedom) as:

$$G(E)_{3D}\,dE = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}dE \tag{3.229}$$

for $E \geq 0$. There are no available states at $E = 0$, and the effective mass of the electron takes into account the effect of the periodic potential on the electron. The minimum energy of the electron is the energy at the bottom of the conduction band (CB), E_c. For the valence band (VB), similar results hold. In the case that the origin of the energies is not chosen to be the bottom of the band (i.e., $E_c \neq 0$), for the conduction band Equation 3.229 comes with an energy term $(E - E_c)^{1/2}$, and in the case of the valence band the energy term is given as $(E_v - E)^{1/2}$. Here the contributions from the light- and heavy-hole bands add to give the total DOS. Note that, just like the DOS in a 3D ideal gas (Equation 3.19), we obtain a square root dependence in energy E. This function is illustrated in Figure 3.119 (full line). Thus, in the case of free 3D motion, the electronic DOS has a smooth square root dependence on energy E with three continuously varying wave vectors $\mathbf{k}_x$, $\mathbf{k}_y$, and $\mathbf{k}_z$. This holds for materials that are large compared with the de Broglie wavelength (L_1, L_2, $L_3 > \lambda_{DB}$). An important feature of nanostructure devices is that their DOS is very different from this expression, and

G(E)dE =

$\frac{1}{V}$ × [the number of states in the energy range from E to E + dE]

G(E)f(E)dE =

$\frac{1}{V}$ × [the number of filled states in the energy range from E to E + dE]

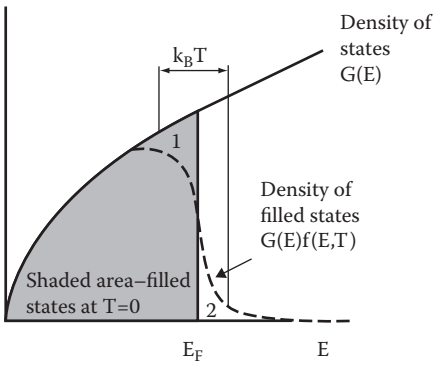


FIGURE 3.119 Density of states (DOS) (full line) and occupied states (shaded area filled at T = 0 K) around the Fermi energy [$G(E)_{3D}$]. The density of filled states is marked with a broken line. As the temperature increases above T = 0 K, electrons from region 1 are excited into region 2.

this will turn out to have important consequences for their electrical and optical properties.

Equation 3.229 gives us the number of possible electronic states of a large 3D device (L_1, L_2, $L_3 > \lambda_{DB}$). To deduce how these states are occupied with electrons, we need to multiply the density of state function [$G(E)_{3D}$] with the Fermi-Dirac distribution function, $f_{FD}(E, T)$, which gives the probability that a state of energy E is occupied by an electron. The number of electrons with energies between E and $E + dE$ is then written as:

$$n(E)_{3D}dE = G(E)f_{FD}(E)dE \tag{3.230}$$

where $G(E)_{3D}dE$ is the number of states between E and $E + dE$. This function is shown as a broken line in Figure 3.119.

The integral of $n(E)_{3D}dE$ over all energies gives the total number of electrons n_{3D} per unit volume (=n_{3D}/V). From Equation 3.230, with f(E) = 1 (T = 0 K), the number density of filled states n_{3D} is:

$$n_{3D} = \int_0^{\infty} n(E)_{3D}dE = \int_0^{E_F} G(E)_{3D}f(E)dE \text{ or}$$
$$n_{3D} = \int_0^{E_F} \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}dE = \frac{\left(\frac{2m_e^*E_F}{\hbar^2}\right)^{\frac{3}{2}}}{3\pi^2} \tag{3.231}$$

Note that at T = 0, $n(E)_{3D}$ is zero for $E > E_F$, so we only have to integrate from E = 0 to E = E_F. In Figure 3.119 this integral corresponds to the shaded area.

For a metal with a total number n_{3D} of valence electrons per unit volume, we can now calculate the maximum energy (E_F) and the maximum **k** value ($\mathbf{k}_F$). The maximum energy (the energy at the surface of the sphere) at T = 0 K is obtained by solving for E_F:

$$E_F = \frac{\hbar^2}{2m_e^*}\left(3\pi^2 n_{3D}\right)^{\frac{2}{3}} = (0.365\ \text{eV nm}^2)(n_{3D})^{\frac{2}{3}} \tag{3.232}$$

This Fermi level is the top of the collection of electron energy levels at absolute zero temperature, and it depends on the number of electrons per unit volume (n_{3D}). Example numbers for Fermi energies for different metals can be found in Table 3.13.

Because fermions cannot exist in identical energy states (see the exclusion principle), at absolute zero, electrons pack into the lowest available energy states and build up a "Fermi sea" of electron energy states. In this state (0 K), the average energy per electron in a 3D electron gas is calculated as:

$$E_{av} = \frac{\text{Total Energy}}{\text{Number of Electrons}} = \frac{\int_0^{E_F} EG(E)_{3D}dE}{\int_0^{E_F} G(E)_{3D}dE}$$
$$= \frac{1}{n_{3D}}\int_0^{E_F} EG(E)_{3D}dE = \frac{3}{5}E_F \tag{3.233}$$

For Cu, for example, this average energy is 4 eV, huge compared with typical thermal energies of 0.025 eV (kT at 300 K). Contrast this result with that of a gas of molecules where a Boltzmann distribution leads to an energy E of zero at 0 K and of the order of kT at a temperature T!

The total number N of **k**-states within the Fermi sphere can be calculated as:

$$N = \frac{\frac{4}{3}\pi k_F^3}{\left(\frac{2\pi}{L}\right)^3} = \frac{k_F^3}{6\pi^2}v \tag{3.234}$$

because each **k**-state occupies $(2\pi/L)^3$ and the total volume of the Fermi sphere equals $\frac{4}{3}\pi k_F^3$. The ratio of

these two gives the total number of states N, each of which accommodates two electrons; in other words, the density of states $n_{3D}(=N/V)$ at the Fermi level is:

$$n_{3D} = \frac{2N}{V} = \frac{k_F^3}{3\pi^2} = \frac{\left(\frac{2m_e^* E_F}{\hbar^2}\right)^{\frac{3}{2}}}{3\pi^2} \qquad (3.235)$$

giving us back Equation 3.231. Thus, the density of single-particle states available per unit volume per unit energy of states at the Fermi level in an unrestricted (3D) electronic device is derived as:

$$G(E)_{3D}(E = E_F) = \left.\frac{dn_{3D}}{dE}\right|_{E=E_F} = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{3}{2}} E_F^{\frac{1}{2}} \qquad (3.236)$$

This is a very important quantity as the rates of many processes are proportional to $G(E)$. Increasing n_{3D}, the density of electrons in 3D increases both E_F (Equation 3.232) and $G(E)_{3D}$ (Equation 3.236). This is of course the same outcome as obtained in Equation 3.229, the only difference being that we replaced E with E_F here.

The result in Equation 3.236 can be simplified by comparing it with Equation 3.235 to obtain:

$$G(E)_{3D}(E = E_F) = \left.\frac{dn_{3D}}{dE}\right|_{E=E_F} = \frac{3n_{3D}}{2E_F} \qquad (3.237)$$

What this means is that the density of single-particle states available per unit volume and per unit energy of states at the Fermi level in 3D is 1.5 times the density of conduction electrons divided by the Fermi energy.

Example 3.7: Calculate the Fermi energy E_F of Na. Na has an atomic density of 2.53×10^{28} atoms/m^3. We are assuming $m_e = m_e^*$, i.e., $h^2/2m_e = 1.505$ eV.nm^2.

Answer: Sodium has one valence electron ($3s$) per atom, so the electron density is $n = 2.53 \times 10^{28}$/m^3. Using Equation 3.232 we obtain:

$$E_F = \frac{1}{4\pi^2}\frac{h^2}{2m_e}\left(3\pi^2 n_e\right)^{\frac{2}{3}}$$

$$= \frac{1}{4\pi^2}\left(1.505 \text{ eV} \cdot \text{nm}^2\right)\left(3\pi^2 \times 25.3 \text{ nm}^{-3}\right)^{\frac{2}{3}} = 3.31 \text{ eV}$$

Obviously electrons in a metal have a very large kinetic energy, even at T = 0!

From Equation 3.235, $\mathbf{k}_F$ is given by:

$$\mathbf{k}_F = (3n\pi^2)^{\frac{1}{3}} \qquad (3.238)$$

which depends only on particle concentration n.

We can now also ask, what is the speed in the highest occupied state E_F? From $\mathbf{k}_F$, the Fermi wave vector, we calculate the velocity $\mathbf{v}_F$ of the electrons on the Fermi surface, i.e., the Fermi velocity (see Equation 3.203, with group velocity $\mathbf{v}_g = \mathbf{v}_F$) as:

$$\mathbf{v}_F = \frac{\mathbf{p}_F}{m_e} = \frac{\hbar \mathbf{k}_F}{m_e} = \sqrt{\frac{2E_F}{m_e^*}} \qquad (3.239)$$

The Fermi velocity is the average velocity of an electron in an atom at absolute zero. This average velocity corresponds to the average energy given above (Equation 3.233). In Equation 3.239, $\mathbf{p}_F$ is the Fermi momentum, i.e., the momentum of fermions at the Fermi surface, and with the expression for $\mathbf{k}_F$ in Equation 3.238 we obtain:

$$\mathbf{v}_F = \frac{\hbar}{m}\left(3\pi^2 n_{3D}\right)^{\frac{1}{3}} \qquad (3.240)$$

We see that in 3D, the higher the electron density, the faster the electrons are moving. In the presence of an electrical field, all the electrons in a conductor move together, so the exclusion principle does not prevent the free electrons in filled states from contributing to the conduction. This is illustrated in Figure 3.120, where we show the Fermi function in one dimension versus velocity at an ordinary temperature. Over a wide range of velocities, the Fermi function equals 1, and speeds $\mathbf{v}_F$ marked in this figure are given by Equations 3.239 and 3.240.

The dashed curve represents the Fermi function after the electric field has been applied for a period t.

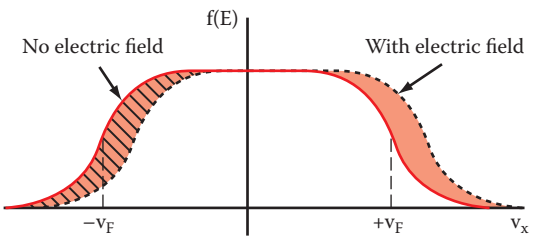


FIGURE 3.120 The Fermi function versus velocity in one dimension for a conductor with (solid) and without (broken line) electrical field in the +x-direction. The effect is greatly exaggerated (see text).

Table 3.12 Fermi Quantities*

Fermi Quantity	Equation 3D	Typical Value
Fermi wave vector	$\mathbf{k}_F = (3\pi^2 n)^{\frac{1}{3}}$	~10^8 cm^{-1}
Fermi energy	$E_F = \frac{\hbar^2}{2m_e^*}(3\pi^2 n)^{\frac{2}{3}}$	~1–10 eV
Fermi wavelength	$\lambda_F = 2\pi/\mathbf{k}_F$	Few nanometers for metals, several tens of nanometers for semiconductors
Fermi temperature	$T_F = E_F/k_B$	~10^4–10^5 K
Fermi momentum	$p = \hbar k_F$	
Fermi velocity	$v_F = \frac{\mathbf{p}}{m_e^*} = \frac{\hbar k_F}{m_e^*}$	~10^8 cm/s
*All values at T = 0 K.		

All the electrons have been shifted to higher velocities, but the net effect is equivalent to shifting electrons near the Fermi level only.

A typical value for the Fermi velocity at 0 K is ~10^6 m/s or 1000 km/s; as pointed out before, this is a surprising result because for a classical gas at 300 K the thermal velocity $\mathbf{v}_{rms} = \sqrt{\frac{8\mathbf{k}_B T}{\pi m}} = 10^5$ m/s (see Equation 3.21) with a velocity that goes to zero at T = 0 K!

The Fermi wavelength λ_F is given as $2\pi/\mathbf{k}_F$; it represents the de Broglie wavelength associated with the Fermi wave vector $\mathbf{k}_F$. The Fermi temperature is the temperature T_F at which $k_B T_F = E_F$. Thus, it is the energy of the Fermi level of an assembly of fermions divided by Boltzmann's constant. The quantity T_F is not to be confused with the temperature of the electron gas. Below the Fermi temperature, a substance gradually expresses more and more quantum effects of cooling. For temperatures much lower than the Fermi temperature, the average energy of the phonons of the lattice will be much less than the Fermi energy, and the electron energy distribution will not differ greatly from that at T = 0. We also recognize here that photons travel much faster than electrons! Photons travel at c = 3.0 × 10^8 m/s (in vacuum) versus electrons that travel at $\mathbf{v}_F$ (Fermi speed) = 1.57 × 10^6 m/s (copper wire).

The Fermi quantities we introduced in this section are summarized in Table 3.12, and in Table 3.13 we list calculated free electron Fermi surface parameters for some metals at room temperature.

Electronic Conductivity for Bulk Materials Drude, at the end of the nineteenth century, could not have known about the Fermi-Dirac distribution. Electrons, being fermions, follow this distribution, and at room temperature it is almost the same as at absolute zero temperature, resulting in a velocity distribution for fermions very different from the one predicted by Maxwell-Boltzmann statistics. Only electrons near the Fermi level contribute to electrical conductivity, and as a result of the wave nature of the electrons, they can pass through a perfect crystal without suffering any resistance at all. This means that the mean free path of an electron passing through a perfect crystal with all nuclei at rest is infinity rather than the interatomic spacing of the order of 1 nm assumed by Drude. In such an ideal crystal, a Bloch function ψ_k evolves into $\psi_{k+\Delta k_x}$, and

Table 3.13 Calculated Free Electron Fermi Surface Parameters for Metals at Room Temperature

Metal/Valency	Electron Concentration n (cm^{-3})	Fermi Wave Vector (cm^{-1})	Fermi Velocity (cm s^{-1})	Fermi Energy E_F (eV)	Fermi Temperature $T_F = E_F/k_B$ (in K)
Cu(1)	8.45×10^{22}	1.36×10^8	1.57×10^8	7.00	8.12×10^4
Ag(1)	5.85	1.20	1.39	5.48	6.36
Au(1)	5.90	1.20	1.39	5.51	6.39
Be(2)	24.2	1.93	2.23	14.14	16.41
Mg(2)	8.60	1.37	1.58	7.13	8.27
Zn(2)	13.10	1.57	1.82	9.39	10.90
Al(3)	18.06	1.75	2.02	11.63	13.49
Ga(3)	15.30	1.65	1.91	10.35	12.01
In(3)	11.49	1.50	1.74	8.60	9.98
Pb(4)	13.20	1.57	1.82	9.37	10.87
Sn(4)	14.48	1.62	1.88	10.03	11.64

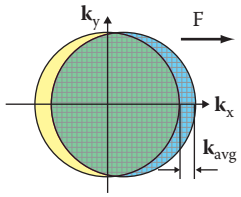


FIGURE 3.121 Displacement of the Fermi surface with an applied electrical field. Collisions with thermal vibrations and defects (not stationary ions or other electrons, as Drude envisaged) stop the Bloch oscillations and cause electrons to settle to a drift velocity. See also Figure 3.71, where the Fermi factor in one dimension for a conductor with (solid) and without (broken line) electrical field in the +*x*-direction is shown.

when this electron state reaches the Brillouin zone at $k = +\pi/a$, it re-enters the crystal at $-\pi/a$ or we get resistance-less Bloch oscillations.

In the absence of an electric field, the same number of electrons is moving in the ±*x*-, ±*y*-, and ±*z*-directions, so the net current is zero. But when a field E is applied, e.g., along the *x*-direction, the Fermi sphere in Figure 3.121, in the absence of collisions, is displaced at a uniform rate by an amount related to the net change in momentum, $\Delta\mathbf{p}_x$, of the free electron gas (FEG) as a whole. The equation of motion (Newton's law) describes this situation as (see Equation 3.208):

$$m_e^* \frac{dv_x}{dt} = \hbar\left(\frac{dk_x}{dt}\right) = -E_x e = F \qquad (3.241)$$

The quantity $\hbar\mathbf{k}$ is the crystal momentum, and thus one can say that the force caused by the electric field is equal to the time derivative of the crystal momentum. Integrating the previous expression we obtain:

$$\mathbf{k}_x(t) - \mathbf{k}_x(0) = -\frac{eE_x t}{\hbar} \qquad (3.242)$$

The shift in Fermi sphere creates a net current flow because more electrons move in the +*x*-direction than the –*x*-direction. According to this model, the Fermi sphere moves with constant velocity in **k**-space. This means that the electron velocity increases indefinitely. This is of course not possible, and it is evident that scattering processes must limit the electron velocity and hence the finite electrical conductivity of metals. A viscous term must be introduced in the equation of motion. In a real crystal, scattering events with a scattering time τ, involving collisions of electrons near the Fermi surface (these are the only ones that can move into empty states), prevent the observation of the ideal Bloch oscillations and oppose the electrical field effect so that the Fermi sphere reaches a steady state when the new center is displaced in the *x*-direction by an average wave vector:

$$\mathbf{k}_{avg} = -\frac{eE_x\tau}{\hbar} \qquad (3.243)$$

The wave vector changes calculated from Equation 3.243 are very small changes; for example, with an electric field *E* of 1 V/m and with a value for the scattering time τ of ~10^{-14} s we calculate a value for $|\mathbf{k}_{avg}|$ of ~15 m^{-1}, a quantity very small compared with a BZ dimension, which we calculate as:

$$\mathbf{k}_{BZ} \approx \frac{2\pi}{a} \approx \frac{2\pi}{3.00\times10^{-10}\,m} \approx 2\times10^{10}\ m^{-1}$$

The effect of the Fermi shift in Figure 3.121 are greatly exaggerated. If at t = 0 an electrical field is applied to a Fermi sphere centered at the origin of **k**-space, the sphere will move to a new position in a characteristic time between scattering events given by:

$$\tau = \frac{\lambda}{v_{avg}} \qquad (3.244)$$

From Equation 3.243 the average velocity of the Fermi sphere is given by:

$$\mathbf{v}_{avg} = \frac{\hbar\mathbf{k}_{avg}}{m_e^*} = -\frac{eE_x\tau}{m_e^*} \qquad (3.245)$$

At the steady state the current density is then given as:

$$\mathbf{J} = -ne\mathbf{v}_{avg} = \left(\frac{ne^2E_x\tau}{m_e^*}\right) = \sigma E \qquad [3.3][3.6]$$

where *n* is the electron density and:

$$\sigma = \frac{ne^2\tau}{m_e^*} \qquad (3.7)$$

These are the same expressions as derived in Drude's model, but with τ given now as $\tau = \frac{\lambda}{v_{avg}}$ (Equation 3.244) and m_e^* replacing m_e. The average velocity,

derived as $\bar{v} = \sqrt{\frac{8k_B T}{\pi m}}$ (Equation 3.21) from the Boltzmann distribution of speeds in Drude's model, is replaced here by $\mathbf{v}_{avg} = \frac{\hbar \mathbf{k}_{avg}}{m_e^*} = -\frac{eE_x\tau}{m_e^*}$ (Equation 3.245). In Drude's model the mean free time between collisions of electrons with lattice ions, τ, is related to the average velocity in $\bar{v} = \mathbf{v}_{dx} = -\frac{eE_x\tau}{m_e}$ (Equation 3.6) as:

$$\tau = \frac{\lambda}{v_x} \approx \frac{a}{v_x} \tag{3.246}$$

Drude assumed the electron mean free path, λ, to be equal to the lattice constant, *a*, which is of the order of 1 nm, in which case this equation yields a typical value for τ of about 10^{-14} s. Based on Equations 3.7, 3.21, and 3.246, Drude then derived the following relationship for the resistivity:

$$\rho = \frac{1}{\sigma} = \frac{m_e \mathbf{v}_d}{ne^2\lambda} = \frac{m_e}{ne^2 a}\sqrt{\frac{8kT}{\pi m_e}} \tag{3.22}$$

Using the lattice constant, *a*, for the mean free path, this equation leads to values for the conductivity of a metal that are six times too small. Moreover, from Equation 3.22 the temperature dependence of the resistivity is determined by v_{dx}, which in this model is proportional to $\sqrt{T}$, whereas in practice the temperature dependence of the resistivity is represented by the empirical relationship:

$$\rho = \rho_0 + \alpha T \tag{3.23}$$

where ρ_0 is the resistivity at a reference temperature, usually room temperature, and α is the temperature coefficient. If the Boltzmann distribution function is applied to the electron gas, one thus immediately finds the velocity of the electron to change as $\sqrt{T}$. According to Drude's model, the mean free path is obviously temperature-independent because it is calculated from the scattering cross-section of rigid ions (with lattice constant *a*). This results in a resistivity proportional to $\sqrt{T}$, provided that the number of electrons per unit volume *n* is temperature-independent (Equation 3.22). However, people at that time had been well aware that the resistivity of typical metals increases linearly with increasing temperature well above room temperature (Equation 3.23). To be consistent with the Maxwell–Boltzmann distribution law, one then had to assume *n* to change as $\frac{1}{\sqrt{T}}$ in metals. This was not physically accepted, and the application of the Maxwell–Boltzmann distribution to the electron system was apparently the source of the problem.

In the quantum mechanical model, only a *few electrons*, all moving at the temperature-independent very high Fermi velocity $\mathbf{v}_F \sim 10^8$ cm/s (Equation 3.239), carry the current, instead of *all electrons* moving at the average drift velocity $\mathbf{v}_d$ (~0.1 cm/s) in Drude's model (see p. 82 "Drude Fails"). Only electrons near the surface of the Fermi sphere find empty orbitals in which they can scatter. Electrons in the inner part of the Fermi sphere find no empty states with similar energy as those electrons near the Fermi sphere. Therefore, inner electrons cannot scatter, and those electrons do not contribute to the current transport process. Replacing v_{th} with v_F in Equation 3.8, one obtains a value for the resistivity that is 100 times larger than the experimental numbers—understandable because v_F is 16 times larger than v_{th}, and we know that the numbers obtained with v_{th} were already six times too large. The resolution lies in the calculation of the mean free path λ; in a perfectly ordered crystal $\lambda = \infty$; in a real crystal, λ is determined by scattering phenomena. We need to replace the inner atomic distance a in Equation 3.22 with the quantum mechanics value for λ, the mean free path of the conduction electrons. Experiments have shown that the electrons can move surprisingly far without any interaction; the mean free electron pass can be up to 10^8 atom distances at low temperatures. Electrons are not scattered by the regular building blocks of the lattice because of the wave character of the electrons. Scattering mechanisms instead are:

1. Lattice defects (foreign atoms, vacancies, interstitial positions, grain boundaries, dislocations, stacking disorders), and
2. Thermal vibration of the lattice (phonons).

Item 1 is more or less independent of temperature, whereas item 2 is independent of lattice defects but dependent on temperature. The mean free path

now does not depend on the radius of the ions but rather on deviations of the ions from a perfectly ordered array such as seen from lattice thermal vibrations and the presence of impurities. The ion vibrations lead to an effective area A that results in an electron mean free path given as $\lambda = 1/n_{ion}A$. Lattice ions are basically points only, but their thermal vibration has them occupy an electron scattering area $A = \pi r^2$, where r is the amplitude of the thermal vibrations. The energy of thermal vibration in a simple harmonic oscillator is proportional to the square of the amplitude of the vibration (r^2). In other words, the area A is proportional to the energy of the vibrating lattice ions. From the equipartition theory we know that the average vibration energy is proportional to kT, so it follows that A is proportional to T and λ is proportional to $1/T$. Because the mean free path is inversely proportional to temperature at high temperatures, it follows that $\sigma \propto \frac{1}{T}$, in agreement with the experimental evidence (Equation 3.23). This solves the issue of the wrong temperature dependence of the resistivity. We need to also calculate a correct absolute value for the resistance using the quantum mechanical mean free path.

The quantum mechanical mean free path of the conduction electrons, say in Cu, is defined as:

$$\lambda = \mathbf{v}_F \tau \tag{3.247}$$

where $\mathbf{v}_F$ is the velocity at the Fermi surface, because all collisions involve only electrons near the Fermi surface. With Fermi velocities $\mathbf{v}_F$ of typically 10^8 cm/s (see Table 3.13) and with room-temperature resistivities for many metals of $\rho \sim 1$–$10\ \mu\Omega\cdot$cm, the corresponding relaxation time is $\sim 10^{-14}$ s and the resulting averaged free electron path at room temperature is about 100 Å. So it is of the order of a few 10s to 100s of interatomic distances. At low temperatures for very pure metals the mean free path can actually be made as high as a few centimeters ($\tau \approx 2 \times 10^{-9}$ s at 4 K for very pure Cu). Compared with using the lattice constant, a, for the mean free path, this equation leads obviously to values for the conductivity that are 100 times higher—in agreement with experimental data.

For a current $\mathbf{J}$ of 1 A/mm^2 in a conductor with an electron density of $n = 10^{22}$ cm^{-3}, we calculate an average speed $\mathbf{v}_{avg}$ of the Fermi sphere of $J/ne \sim 0.1$ cm/s (see Equation 3.3), which is much less than the Fermi velocity $\mathbf{v}_F \sim 10^8$ cm/s. In a conductor, charges are always moving at the Fermi velocity, but the Fermi sphere moves much slower because electrons travel at fast Fermi velocities for a short average time, τ, and then "scatter" because of collisions with atom vibrations, grain boundaries, impurities, or material surfaces (especially in very thin films).

The value for $n(E)_{3D}$ in $\mathbf{J} = -n(E)_{3D}e\mathbf{v}_{avg}$ (Equation 3.3) at $T > 0$ is obtained from Equation 3.230:

$$n(E)_{3D}\,dE = G(E)_{3D} f_{FD}(E)\,dE$$

$$= \frac{8\pi\sqrt{2}m_e^{\frac{3}{2}}}{h^3} E^{\frac{1}{2}} \frac{1}{e^{\frac{E-E_F}{kT}}+1}\,dE \tag{3.248}$$

For a large semiconductor, Equation 3.248 was illustrated in Figure 3.119 with a broken line. We will see below that $G(E)$ depends strongly on dimensionality, so we also expect the current density to vary strongly with dimensionality.

Quantum Wells: $L_{1,2} > \lambda_{DB} > L_3$

Quantization As we saw earlier, devices that come with a length L that in one direction is comparable with the size of the electron de Broglie wavelength are known as quantum wells (1D confinement). A planar quantum well structure may be made from a thin region of a narrow gap semiconductor sandwiched between two layers of a wide bandgap semiconductor. We use for an example of such 1D confinement a quantum well made by sandwiching a layer of GaAs between two layers of $Al_xGa_{1-x}As$, as shown first in Figure 3.59 and, simplified, reproduced in Figure 3.122. Growing two different semiconductors on top of each other, as illustrated,

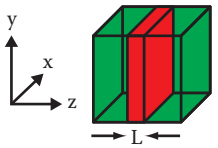


FIGURE 3.122 A quantum well can be made, for example, by sandwiching a layer of GaAs between two macroscopic layers of $Al_xGa_{1-x}As$. This creates a layer in which the electrons (and holes) behave in a 2D way (see also Figure 3.59).

forms heterojunctions. The narrower bandgap GaAs material is enclosed by a material with a considerably larger bandgap to establish a potential barrier at the surface of the confined material. Because of the potential barrier, the motion of electrons and holes is restricted in one dimension (thickness L in the z-direction in this case) and is forced to occupy discrete states of energy instead of staying arbitrarily within an energy continuum. Hence quantization of the system occurs by shrinking the thickness of the GaAs layer. A 2D electron gas in the laboratory is really a 3D electron system in which the electron motion is strongly confined in one spatial direction but in which free motion is still allowed in the other two directions. At the interface of the semiconductors GaAs and $Al_xGa_{1-x}As$ a potential well is formed. The potential well is a result of charge transfer between the two materials and their conduction band offset (E_c), confining the electrons, and the motion of electrons perpendicular to the plane of the heterointerface (z-axis in Figure 3.122) is quantized.

The composition x of the ternary semiconductor $Al_xGa_{1-x}As$ can be varied to control the electron barrier height. A good lattice match between GaAs and $Al_xGa_{1-x}As$ over a wide range of x values minimizes lattice strain at the interfaces. When the electron and holes are confined in the same layer, one talks about a type I quantum well; with electron and holes confined in different layers, one defines a type II quantum well.

For the quantum well considered here, the $Al_xGa_{1-x}As$ layers are thick enough so that tunneling through these layers remains very limited.

For an energy E smaller than the potential barrier V, the energy of an electron in the conduction band of a quantum well is given as:

$$E_n(k_x, k_y) = E_c^0 + \underbrace{\frac{n_z^2 h^2}{8m_e^* L_z^2}}_{E_{1D}(n_z)} + \underbrace{\frac{h^2(k_x^2 + k_y^2)}{8m_e^*}}_{E_{1D}(k_x, k_y)}, \quad n = 0, 1, 2, \ldots \tag{3.249}$$

with the wave function:

$$\psi_n(x, y, z) = \psi_n(z) e^{ik_x x} e^{ik_y y} \tag{3.250}$$

There is one quantized component in the z-direction and a "free" electron component in the x-y plane. The second term on the right side of Equation 3.249 represents the quantized energy in the z-direction (the thickness direction L of the GaAs film). This is the same expression we derived in Equation 3.156 for a finite-sized 1D box with infinitely high potential walls. For quantization to be important, the difference between the electron energy levels should be much larger than the thermal energy k_BT, that is, $E_n = \frac{n_z^2 h^2}{8m_e^* L_z^2} >> k_B T$, where n_z = 1, 2, 3 are the quantum numbers labeling.the energy levels. E_c^0 is the energy corresponding to the bottom of the conduction band. Strictly speaking, the above expressions apply only to an infinitely deep potential well. However, we can use the same equations as long as E_n is well below the bottom of the conduction band of the wide band material. Using this condition, we find, for example, that in GaAs where $m_e^*/m_e = 0.067$, the levels are quantized at room temperature when L_z = 150 Å. The third term on the right side of Equation 3.249 represents the kinetic energy of the electrons in the x-y plane where they are free to move. The $\mathbf{k}_z$-component is absent in the last term of Equation 3.249 because the motion in this direction is quantized. Equation 3.249 reveals that for each value of the quantum number n, the values of wave vector components $\mathbf{k}_x$ and $\mathbf{k}_y$ form a 2D band structure. The wave vector $\mathbf{k}_z$ in the z-direction, on the other hand, can only take on discrete values, $\mathbf{k}_z = n_z\pi/L_z$. For each value of n there is a sub-band with n the sub-band index as illustrated in Figure 3.123. In this figure we show energy levels (bottoms of

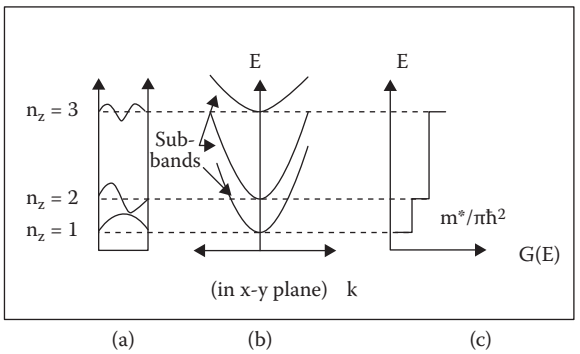


FIGURE 3.123 (a) Energy levels (bottoms of sub-bands) for a quantum well made by sandwiching a layer of GaAs between two macroscopic layers of $Al_xGa_{1-x}As$ (L_z is 150 Å). (b) Energy versus $\mathbf{k} = (\mathbf{k}_x^2 + \mathbf{k}_y^2)^{1/2}$ for 2D electron gas in GaAs quantum well. (c) Density of states for quantum well structure (Harris, 2006.[4])

sub-bands), density of states for a quantum well structure, and energy versus $\mathbf{k} = (\mathbf{k}_x^2 + \mathbf{k}_y^2)^{1/2}$ for the 2D electron gas in the GaAs quantum well.

We can carry out an analogous argument for the holes in the valence band with the difference that their quantized energy is inverted and that we need to invoke m_h^* for the effective mass of the hole. For a quantum well, the lowest-energy band-to-band (interband) transition is now different from the bandgap (E_g) transition of the bulk semiconductor. It will occur at a higher energy level (shorter wavelength) between the lowest energy state for electrons in the conduction band (n = 1) and the corresponding state for holes in the valence band. This defines the effective bandgap for a quantum well as:

$$E_g^* = E_c^0 - E_v^0 + \frac{h^2}{8L^2}\left(\frac{1}{m_e^*} + \frac{1}{m_h^*}\right) \qquad (3.251)$$

The shift to higher wavelengths is referred to as a "blue shift," caused by quantization.

Fermi Surfaces and Brillouin Zone for Quantum Wells In the case of a free 3D electron gas we appreciate that the surface of the Fermi sea is a sphere of radius $\mathbf{k}_F$ connecting points of equal energy in **k**-space (see Figure 3.111). In 2D this becomes a circle connecting points of equal energy in 2D **k**-space. In Figure 3.124 we show the Fermi circles corresponding to 2D crystals with one, two, three, and four valence electrons per atom. In this figure we also show the 2D Brillouin zone (see square in dashed line), and we see how the free electron circle of a three-valent metal (red circle) cuts the Bragg planes located at π/a. The square shown in Figure 3.124 corresponds to the first Brillouin zone of a 2D square lattice (Figure 3.112b).

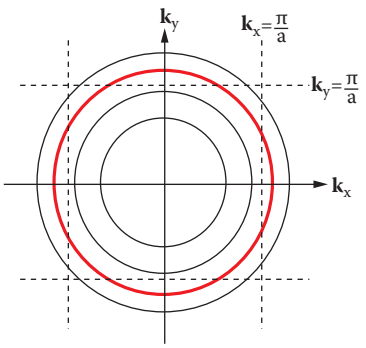


FIGURE 3.124 Free electron circles for 1, 2, 3, and 4 valence electrons. The free electron circle of a three-valent metal (red circle) cuts the Bragg planes located at π/a.

Density of States for Quantum Wells The density of states for each sub-band of a quantum well (Figure 3.123c) can be found using an approach similar to the one we used above for a 3D density of states function, that is, by counting the number of states with wave vectors **k** between **k** and $d\mathbf{k}$. In the case of 1D confinement of electrons we must find the number of **k**-states enclosed in an annulus of radius $\mathbf{k} + d\mathbf{k}$ (see Figure 3.125). Each state occupies an area A of $\left(\frac{2\pi}{L}\right)^2$. The area of the annulus, A, is given by:

$$A d\mathbf{k} = 2\pi k d\mathbf{k} \qquad (3.252)$$

Dividing the area of the annulus by the area occupied by a **k**-state, and remembering again to multiply by 2 for the electron spin states, we get for the number of states per unit area:

$$G(\mathbf{k})_{2D}\, d\mathbf{k} = 2\frac{2\pi k d\mathbf{k}}{\left(\frac{2\pi}{L}\right)^2} = \frac{L^2\mathbf{k}}{\pi} d\mathbf{k} \qquad (3.253)$$

Or, in terms of energy per unit area at an energy E (dividing by L^2), the density of states for each sub-band is given as:

$$G(E)_{2D} dE = \frac{k d\mathbf{k}}{\pi} = \sqrt{\frac{2m_e^* E}{\hbar^2}}\left(\frac{2m_e^* E}{\hbar^2}\right)^{-\frac{1}{2}} \frac{m_e^*}{\pi\hbar^2} dE$$
$$= \frac{m_e^*}{\pi\hbar^2} dE \qquad (3.254)$$

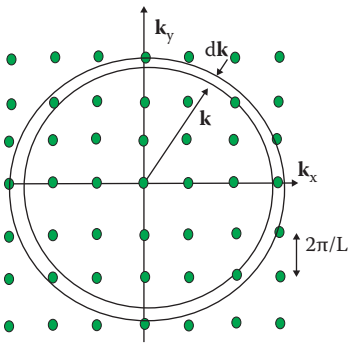


FIGURE 3.125 Density of states, $G(\mathbf{k})_{2D}$, is the number of allowed states between **k** and $\mathbf{k} + d\mathbf{k}$, i.e., the number of states between a sphere of radius **k** and a sphere of radius $\mathbf{k} + d\mathbf{k}$.

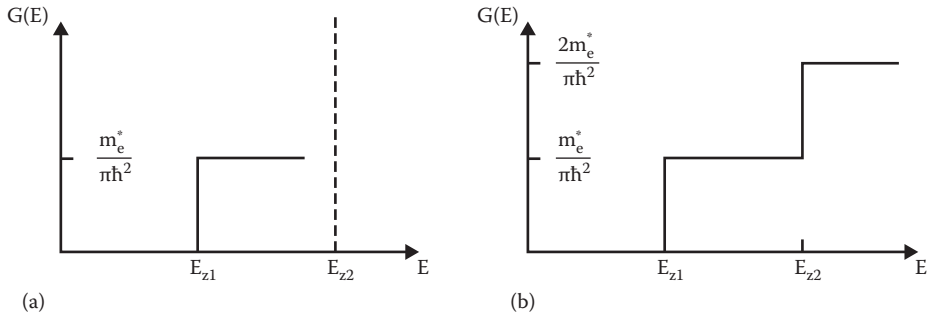


FIGURE 3.126 Filling of the first two sub-bands in a 2D structure. The first sub-band has a constant energy at $\frac{m_e^*}{\pi\hbar^2}$ (a), and the second has a constant energy at $\frac{2m_e^*}{\pi\hbar^2}$ (b).[4]

For this derivation we used Equation 3.227, which holds true for any dimensionality D of the problem. Importantly, in an ideal 2D system, the density of states is constant and does not depend on energy.

The density of states in a 2D electron gas, in which only the lowest energy sub-band (n = 1 level E_{z1}) is occupied, is illustrated in Figure 3.126a, where we assume that the confinement is in the z-direction. For all electron energies up to E_{z1} the density of states is zero because electrons cannot exist in the well at lower energies than this. Because of the freedom of motion in the plane of the hetero interface (with continuously varying wave vectors $\mathbf{k}_x$ and $\mathbf{k}_y$ in the x-y plane), the energy levels that form in the potential well are highly degenerate. If the electron density is sufficiently low, then all electrons can be accommodated in the lowest level of the well, and the freedom of motion in the transverse direction is frozen out and the electron system effectively behaves as the ideal 2D one as described by Equation 3.254 with a single step in energy only with a constant value $\frac{m_e^*}{\pi\hbar^2}$. In the regime where only the lowest sub-band is occupied, increasing energy corresponds to increasing electron motion in the x-y plane (kinetic energy).

As the electron filling of the quantum well is increased, eventually electrons begin to fill the next transverse level of the potential well (Figure 3.126b). At this point we have two so-called 2D sub-bands occupied in the quantum well. To a good approximation, electrons in the two sub-bands may be viewed as forming two independent 2D electron gas systems. Thus, the density of states is double that which we would expect in the case where just a single sub-band is occupied. For each quantum state in the quantum well, there will be a step in the density of states. The overall density of states is discontinuous, with a stepwise structure that is characteristic of quantum wells. Because of the summing over the different sub-bands, a more general description of a 2D system has Equation 3.254 modified as:

$$G(E)_{2D} = \frac{m_e^*}{\pi\hbar^2}\sum_n H(E - E_n) \qquad (3.255)$$

where $H(E - E_c)$ is the Heaviside step function.* It takes the value of zero when E is less than E_n and 1 when E is equal to or greater than E_n. E_n is the nth energy level within the quantum well.

In Figure 3.127 we summarize the characteristics of a 2D density of states function $G(E)_{2D}$. The energy spacing shown here increases with decreasing L; the thinner the 2D film, the more it approximates an ideal 2D gas with only one sub-band.

The step-like behavior of a 2D density of states function $G(E)_{2D}$ implies that the density of states in the vicinity of the bandgap is much larger than in the case of a bulk semiconductor, where the value of $G(E)_{3D}$ goes to zero (E_1 versus E_g). This makes for stronger optical transitions because a major factor

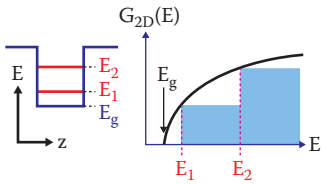


FIGURE 3.127 The density of states function for a quantum well. The solid black curve is that of a free electron. The bottom of the quantum well is at energy E_g.

* The Heaviside step function, H, also called the unit step function, is a discontinuous function whose value is zero for negative argument and one for positive argument.

in the expression for the probability of optical transitions is the density of states. The strength of an optical transition is often defined as the oscillator strength, and the oscillator strength of a quantum well in the vicinity of the bandgap is considerably enhanced compared with a bulk semiconductor. Below we will show how Fermi's golden rule for absorption describes transition rates between levels in terms of the availability of states [density of states: $G(E)$], the availability of photons (intensity E_0), and a "coupling strength" between the levels (transition matrix element $\left|H'_{v,c}\right|^2$). The enhanced oscillator strength of quantum-confined structures is put to good use in the fabrication of laser media for highly efficient and compact solid-state lasers (see Chapter 5). In Figure 3.128 we illustrate the bound-state energies for electrons in the conduction band and for holes in the valence band for a GaAs/$Al_xGa_{1-x}As$ heterojunction. If the GaAs layer is thin enough, bound states form as indicated here by dashed lines. For a perfect 2D electron gas, each bound state corresponds to a discontinuous jump in the electronic density of states function. The density of states function (DOS) can be investigated by measuring the absorption α of electromagnetic radiation associated with the excitation of an electron from the valence band to the conduction band. Because the wavelength of the radiation is long compared with the width of the well, transitions only occur between states for which the spatial variation of the wave function is similar; this leads to the selection rule $\Delta n = 0$ for the adsorption. Therefore, the allowed transitions are those indicated by the arrows in Figure 3.128a. The frequency dependence of the absorption should reflect the steps in the DOS with the steps in the absorption expected to occur at frequencies ω_n given by:

$$\hbar\omega_n = E_{Cn} - E_{Vn} \tag{3.256}$$

where $E_{Cn} - E_{Vn}$ is the energy difference between the nth bound states in the conduction and valence bands. The measured adsorption spectra for GaAs layers of thickness 140, 210, and 4000 Å are shown in Figure 3.128b, and the expected step structure is clearly visible for the two thinner layers, with arrows indicating the frequencies at which steps are expected. Peaks at energies just a little below the predicted values mark the absorption. These results from the creation of an exciton, an electron-hole bound state that is created when a photon is absorbed (see below for details). Because there is an attraction between the electron and the hole in an exciton, the photon energy required to create an exciton is lower than the predicted values that ignore such interactions. Thus, the difference in energy of the peak and the predicted absorption edge is a measure of the binding energy of the electron-hole pair. An exciton, with an energy just below the bandgap, is clearly seen in the absorption curve for the 4000-Å layers, but the step-like structure has disappeared, indicating that the DOS is a smooth curve as one expects for 3D behavior.

Electronic Conductivity of Quantum Wells The circles shown in Figure 3.124 are the Fermi circles corresponding to 2D crystals with one, two, three, and

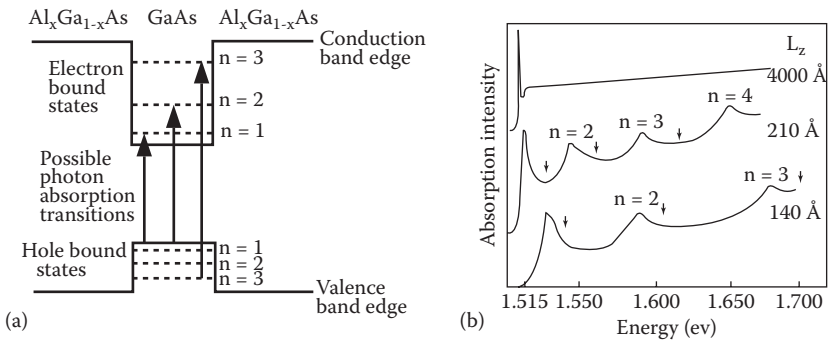


FIGURE 3.128 (a) Heterojunction of GaAs and $Al_xGa_{1-x}As$. If the GaAs layer is thin enough, bound states (dashed lines) form. When photons are absorbed, electrons are excited between these bound states. (b) The adsorption of light, measured as a function of photon energy for GaAs layers of thickness 4000, 210, and 140 Å. The arrows indicate the energies at which the onset of adsorption is expected to occur for transitions involving the nth bound state.

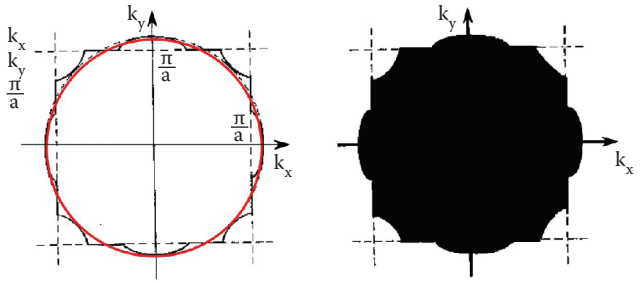


FIGURE 3.129 Deformation of the free electron circle near Bragg planes, $V(x) \neq 0$.

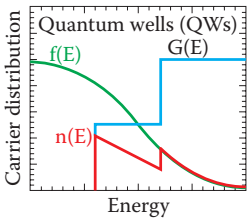


FIGURE 3.130 Quantum wells (QWs). Density of states (DOS) (blue) and occupied states (red) around the Fermi energy for a quantum well. The Fermi-Dirac function is $f(E)$, and the product of $f(E)G(E)_{2D} = n(E)_{2D}$ at $T > 0$.

four valence electrons per atom. The circles represent surfaces of constant energy for free electrons [$V_{(x)} = 0$], i.e., the Fermi surface for some particular value of the electron concentration. The total area of the filled region in **k**-space depends only on the electron concentration and is independent of any interaction of electrons with a lattice. In a solid more realistic shape of the Fermi surface depends on the lattice interaction of the electrons and will usually not be an exact circle in a lattice. There is a discontinuity introduced into a free electron Fermi circle any time it approaches a 2D Brillouin zone boundary. In Figure 3.129 we see how the shape of the Fermi circle is distorted near the surface if $V(x)$ is not zero.

Energy gaps appear at zone boundaries, and the Fermi surface intersects zone boundaries almost always perpendicularly. The crystal potential causes rounding of sharp corners on a Fermi surface. The volume enclosed by a Fermi surface only depends on electron density and not on details of the lattice interaction. In other words, the volume enclosed by a Fermi surface remains unchanged under the "deformations" just mentioned.

The number of occupied states per unit volume in the energy range E to $E + dE$ and with $f(E) = 1$ $(T = 0)$ is calculated as:

$$n_{2D} = \int_0^{\infty} n(E)_{2D} dE = \int_0^{E_F} G(E)_{2D}[f(E) = 1]dE \tag{3.257}$$

or

$$n_{2D} = \int_0^{E_F} \frac{m_e^*}{\pi\hbar^2} dE = \frac{m_e^* E_F}{\pi\hbar^2}$$

To calculate the current density for a 2D gas at a particular temperature we substitute the value for $n(E)_{2D}$ at $T > 0$ [and thus $f(T)$ can then be different from 1] in $\mathbf{J} = -n(E)_{2D}e\mathbf{v}_{avg}$ (Equation 3.3). The function $n(E)_{2D}$ for a given temperature T (>0) is shown as a red line in Figure 3.130. In the same graph we also show the Fermi-Dirac function and the DOS function (blue).

The form of the density of states function of a 2D gas can also be dramatically modified by a magnetic field, giving rise to very pronounced behavior in the conductance with the appearance of the so-called Landau levels.

Quantum Wires: $L_1 > \lambda_{DB} > L_{2,3}$

Quantization We saw earlier that when L becomes very small along two directions (2D confinement)—of the order of the de Broglie wavelength of an electron—one obtains a quantum wire where electrons can only move freely in one direction, i.e., along the length of the quantum wire as shown first in Figure 3.63 and, simplified, reproduced in Figure 3.131. These 1D electronic structures with 2D quantum confinement comprise nanowires, quantum wires, nanorods, and nanotubes.

The starting point for the fabrication of one type of quantum wire is a 2D electron gas confined in one direction as discussed above. A 2D quantum gas that is very strongly confined at some interface and where we can assume that only the lowest subband of the electron gas is occupied, so that the motion transverse to that interface is frozen out. This is the situation we encountered for a

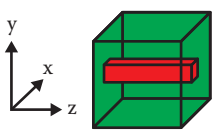


FIGURE 3.131 A rectangular quantum wire.

heterojunction with a very small L, so that Equation 3.254 is applicable [$G(E)_{2D} = \frac{m_e^*}{\pi\hbar^2}$ (no summation of steps)]. To this strong confinement, a typically weaker, lateral confinement of the electrons is added by etching nanowires in the 2D quantum well. In Figure 3.131 we drew a rectangular quantum wire with a square cross-section; in reality nanowires will typically be much wider than they are thick (the x-direction is assumed to be the film thickness direction), so the quantum confinement is most severe in the x-direction. The reason for this discrepancy is that in micro- and nanotechnology it is much easier to control a film's thickness (x-direction) than it is possible to control the lateral dimensions of a structure (y direction). Let us consider the 75-nm-wide quantum wires etched in a $GaAs/Al_xGa_{1-x}As$ heterojunction in Figure 3.132. The 2D electron gas formed at the heterointerface is confined here in a scale of just a few nanometers (in the thickness or x-direction), and so its quantized energies are large in the x direction. The lateral confinement (y direction) of electrons in the 75-nm-wide wire is much weaker than this and its quantized energies are consequently much smaller. The weaker lateral confinement in the wires gives rise to a series of relatively closely spaced energy levels (see E_Y in Figure 3.132). Therefore, transport through the wire in the z-direction will involve electrons that occupy many of these lateral sub-bands. Thus, these structures in reality are quasi-2D confined systems with free electron motion in one direction and two different types of confinement in the other two.

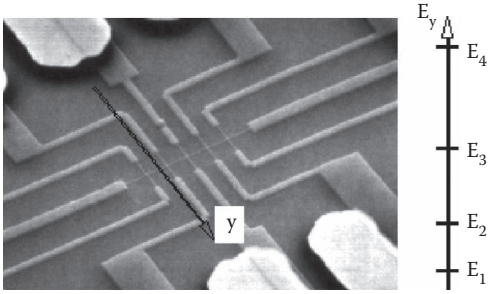


FIGURE 3.132 75-nm-wide quantum wires etched in a $GaAs/Al_xGa_{1-x}As$ heterojunction. The confinement of electrons in the x- and y-directions quantizes the electron energy into a set of discrete energies. The confinement in the x-direction is much stronger than in the y-direction.

For an idealized nanowire with a square cross section and the x = y dimensions in the nanoscale but continuous along the wire axis (z-direction), the energy dispersion function may be written as:

$$E = E_{n_1,n_2}(k_z) + \frac{\hbar^2 k_z^2}{2m_e^*} \qquad (3.258)$$

with the wave function:

$$\psi(x, y, z) = \psi_{n_1,n_2}(x, y)e^{ik_{yz}} \qquad (3.259)$$

with n_1 and n_2 the quantum numbers labeling the eigenstates in the x-y plane and $\mathbf{k}_z$ the wave vector in the z-direction, and where we assume that the dispersion relation for the electron energy is parabolic. The energy term has again two contributions: one is caused by the continuous band value $\left(\frac{\hbar^2 k_z^2}{2m_e^*}\right)$, and the other term [$E = E_{n_1, n_2}(k_z)$] is that of the quantized values of electrons confined in two dimensions as derived in Equation 3.158 (unconstrained direction is along the z-axis). Referencing energies to the conduction band edge, the latter expression may be rewritten as:

$$E_{n_1,n_2}(\mathbf{k}_z) = E_c^0 + \frac{h^2}{8m_e^*}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right) + \frac{\hbar^2 k_z^2}{2m_e^*} \qquad (3.260)$$

where E_c^0 is the bottom of the conduction band. The lowest sub-band is obtained for n_1 and $n_2 = 1$. The energy at the bottom of each sub-band ($\mathbf{k}_z = 0$) is simply given by:

$$E_{n_1,n_2} = E_c^0 + \frac{h^2}{8m_e^*}\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right) \qquad (3.261)$$

If we choose the cross-section of the GaAs quantum wire containing the 2D confined electron gas to be equal to 100 × 100 Å, then the lowest energies in the two lowest sub-bands are equal to 0.112 eV and 0.280 eV. This is determined from Equation 3.261 for $n_1 = 1$, $n_2 = 1$ and $n_1 = 1$, $n_2 = 2$, respectively. In the case that the nanowire is much wider (y-direction) than it is thick (x-direction), the quantum confinement is most severe in the x-direction. Because in this case $L_y >> L_x$, the n_2 levels form a staircase of small steps in the widely separated sub-bands corresponding to the various values for n_1. Thus, confinement that is different in the x and y directions splits each

sub-band further up into a set of more narrowly spaced sub-bands.

In Figure 3.133 we show band structure and density of states for a quantum wire with a square cross section; similar to the situation in quantum wells, sub-bands are formed, but in quantum wires with a rectangular cross-section multiple sub-bands form at each eigenvalue $E(n_x,n_y)$, spread out as $\frac{\hbar^2 k_z^2}{2m_e^*}$.

Fermi Surface of Quantum Wires Whereas Fermi surfaces in 3D and 2D electronic structures consist of a sphere and a circle, respectively, the Fermi surface of a strictly 1D electronic structure consists of just two points at $+\mathbf{k}_F$ and $-\mathbf{k}_F$ or $\int d\mathbf{k} = 2\mathbf{k}_F$ (where $\mathbf{k}_F$ is the Fermi wave vector) or $n_{1D} = \frac{2\mathbf{k}_F}{\pi}$. This unusual Fermi surface has some pretty significant consequences. In 1981, Hiroyuki Sakaki predicted that ideal 1D electrons moving at the Fermi level in quantum wires would require very large momentum changes ($\Delta k = 2\mathbf{k}_F$) to undergo any scattering. The result is that electron scattering is strongly forbidden. This is a consequence of the fact that in one dimension, electrons can scatter only in one of two directions: forward and 180° backward. With this large reduction in scattering, electrons should achieve excellent transport properties (e.g., very high mobility).

Density of States Function for Quantum Wires For calculating the density of states for a quantum wire, we use the same approach we took for the 3D and 2D cases. The **k**-state has now a length l of $2\pi/L$, and we must find the number of **k**-states lying in a length of $\mathbf{k} + d\mathbf{k}$. The wire length difference, l, is given by:

$$l = 2 \tag{3.262}$$

The factor of two appears because the wave number could be either positive or negative, corresponding to the two directions along the wire.

The resulting density of states per unit length of 1D **k**-space is obtained by multiplying by 2 for spin degeneracy and dividing by $2\pi/L$:

$$G(\mathbf{k})_{1D}\,d\mathbf{k} = 2\left(\frac{2}{2\pi/L}\right)d\mathbf{k} = \frac{2L}{\pi}d\mathbf{k} \tag{3.263}$$

To obtain the density of states in terms of the energy $[G(E)_{1D}]$, we use again the relation between E and **k**, derived earlier as $\mathbf{k} = \left(\frac{2m_e^* E}{\hbar^2}\right)^{\frac{1}{2}}$ (Equation 3.150) and differentiate the latter expression with respect to energy:

$$d\mathbf{k} = \left(\frac{2m_e^* E}{\hbar^2}\right)^{-\frac{1}{2}} \frac{m_e^*}{\hbar^2} dE \tag{3.227}$$

The density of energy states $G(E)_{1D}$, the number of allowed states between E and $E + dE$, per unit energy and unit length (divide by L) is then obtained using the chain rule:

$$G(E)_{1D}dE = \frac{2d\mathbf{k}}{\pi} = \frac{1}{\pi}\left(\frac{2m_e^*}{\hbar^2}\right)^{\frac{1}{2}} \frac{1}{E^{\frac{1}{2}}} dE = \frac{1}{h\pi}\sqrt{\frac{2m_e^*}{E}}dE \tag{3.264}$$

This is in sharp contrast with the behavior of a 3D electron gas where $G(E)_{3D}$ goes to zero at low energies, and two dimensions, where $G(E)_{2D}$ steps up to a constant value at the bottom of each 2D sub-band. Remembering that the group velocity is given as:

$$\mathbf{v}_g = \frac{d\omega}{d\mathbf{k}} \tag{3.101}$$

which we found may also be written as:

$$\mathbf{v}_g = \frac{1}{\hbar}\left[\frac{dE}{d\mathbf{k}}\right] \text{ or more generally } \mathbf{v}_g = \frac{1}{\hbar}\nabla_k E(\mathbf{k}) \tag{3.202}$$

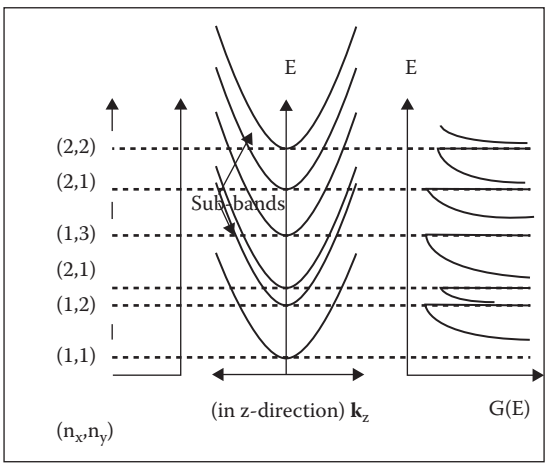


FIGURE 3.133 Band structure and density of states for a quantum wire with a square cross-section.[4]

we can rewrite Equation 3.264 now also in terms of the group velocity as:

$$G(E)_{1D}dE = \frac{2dk}{\pi} = \frac{2}{\pi}\left(\frac{dE}{dk}\right)^{-1} dE = \frac{2}{\pi\hbar v_g(E)}dE \quad (3.265)$$

Thus, for a 1D system the density of states is inversely proportional to the velocity! We will use this expression in v_g when we derive the expression for the current through a nanowire. We will learn that in a nanowire the current is constant and proportional to the velocity and density of states (see further below Equation 3.269). If we add electrons to the nanowire, they initially fill only the lowest of the levels, say level (1,1) in Figure 3.133. As the energy is increased and remains less than level (1,2), the lateral motion remains frozen, and the increase in energy is transferred into motion along the length of the wire. The density of states in this level will then take the form predicted by Equation 3.265. However, in most situations the Fermi energy of the electrons in the wire is several times larger than the average spacing between the lateral energy levels in Figure 3.133, so several of these levels will be occupied, and each of these levels defines a corresponding 1D sub-band. The density of states within each sub-band is 1D, but the total density of states is obtained by summing over all sub-bands n. The summing over the individual sub-bands can be formulated mathematically as:

$$G(E)_{1D} = \frac{1}{\pi}\left(\frac{m_e^*}{\hbar^2}\right)^{\frac{1}{2}} \sum_n L\left(\frac{1}{E - E_n}\right)^{\frac{1}{2}} \quad (3.266)$$

L is the unit step function (same as H in Equation 3.255). The density of states of a quantum wire diverges as $\left(\frac{1}{E - E_n}\right)^{\frac{1}{2}}$ at each sub-band threshold and has the inverse energy dependence $E^{-1/2}$ compared with the $E^{1/2}$ dependence of a 3D electron gas (bulk semiconductor). The DOS of a quantum wire obviously has a more pronounced structure than does a 2D well, with a large number of sub-bands, each one starting as a peak. An immediate manifestation of a large $G(E)_{1D}$ at the bottom of each sub-band is again an increase of the strength of optical transitions or oscillator strength as compared with 3D and 2D electronic structures.

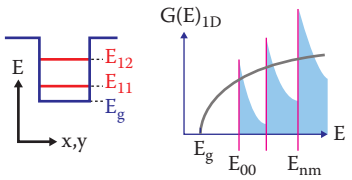


FIGURE 3.134 The density of states function for a rectangular quantum wire. The solid black curve is that of a free electron.

In Figure 3.134 we summarize the characteristics of a 1D density of state function $G(E)_{1D}$. The axis of the rectangular wire is again in the z-direction, and quantization is in the x- and y-directions. The peaks show the calculated density of states of a quantum wire over a range of energies where several different sub-bands become occupied. For comparison, the solid line shows the monotonic variation of the density of states expected for a 3D system.

Electronic Conductivity of Quantum Wires Ohm's law for macroscopic systems is given by V = IR, and in terms of conductance σ this is equivalent to σ = J/E (Equation 2.2). From the previous sections, we know that only electrons close to the Fermi level contribute to the conductance. From Bloch's theorem, we also recall that in an ideal periodic potential, electrons propagate without any scattering—and thus no resistance at all—but that electron propagation in real materials does involve scattering. The origin of such scattering can be any source of disorder that disturbs the perfect symmetry of the lattice. Examples include defects and impurities, scattering from other electrons, and lattice vibrations (phonons). Because an electric current constitutes a movement of those "bumbling" electrons under an electric field, one expects that with nanostructures featuring dimensions comparable with the fundamental size of the electron, electrical properties will be strongly influenced by quantum-mechanical transport effects.

Each of the discrete peaks in the density of states (DOS) in Figure 3.134 is caused by the filling of a new lateral sub-band. The peaks in the density of states functions at those energies where the different sub-bands begin to fill are called criticalities or Van Hove singularities. These singularities (sharp peaks) in the density of states function lead to sharp