Manijeh Razeghi

Fundamentals of Solid State Engineering

Fourth Edition

EXTRAS ONLINE

Springer

Fundamentals of Solid State Engineering

Manijeh Razeghi

Fundamentals of Solid State Engineering

Fourth Edition

Springer

Manijeh Razeghi
Department of Electrical Engineering & Computer Science
Northwestern University
Evanston, IL, USA

Additional material for this book can be downloaded from
https://www.springer.com/us/book/9783319757070

ISBN 978-3-319-75707-0 ISBN 978-3-319-75708-7 (eBook)
https://doi.org/10.1007/978-3-319-75708-7

Library of Congress Control Number: 2018943682

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Students commonly think of a textbook as merely a tool to get prepared for exams. This is not the right way of looking at it! A textbook is the fruit of long-term studies and experience acquired by the author and reflects her or his personality. It embodies priorities, knowledge, and I dare say even dreams and life attitudes. Compare the difference in style and content in the now classic physics textbooks by Landau and Feynman. Both Landau and Feynman were scientists whose minds were ready to listen to the music of the heavens. But how very differently! Landau wrote with the authority of a Zeus and his book sounds like the ultimate message from Heaven, while Feynman's style is more modest, and his curiosity and quest for truth could hardly be matched by anyone. His famous textbook is like an invitation to travel through the Disneyland of Nature, where he acts as a guide, but a guide who is also learning during this journey. And there is a third example: the Chicago lecture notes on quantum mechanics by another Nobel laureate – Enrico Fermi. At first sight, it appears to be more student friendly, simple, and very much to the point, but what a simplistic and, indeed, incorrect interpretation that would be! Fermi made a selection of topics and then reduced the content to the absolute essence of what has to be understood to get prepared for a journey into the quantum wonderland. He did it in such a way that an average student had the impression he or she understood everything, while a more demanding student would get a sense of much more: a feeling that a miraculous quantum world was waiting for him behind invisible doors, full of questions and surprises. Fermi did what Albert Einstein once said about science in his peculiar English – *do it simple, but not simpler*.

I admire this textbook by Professor Razeghi as much as I respect her research achievements, which she fulfilled in her personal journey through this demanding life. She was born in Persia, but left her motherland forever to join her new country France, the country that gave her the chance to continue the science she loved so much. In doing so, she followed the footsteps of Marie Curie, who a century before left oppressed Poland as a young math teacher by the name of Skłodowska. Welcomed in France, Skłodowska completed her studies at the Sorbonne, got married to a brilliant French physicist Pierre Curie, and then spent endless hours working with him, processing tons of radioactive ores from Czechoslovakia. Together they eventually extracted small grains of the miraculous polonium and radium – two radioactive elements they discovered and named. This superb

technological achievement, of which Marie definitely was the master and the *spiritus movens*, opened new avenues for science and finally led her twice to Stockholm to be awarded the Nobel medal.

Dr. Razeghi hopefully was not forced to work in a cold and primitive warehouse, like the Curies had to. The wise management of the French electronic giant Thomson spotted her unique talents and gave her proper resources to realize her visions and dreams. In a short time she became the First Lady in solid-state physics and made Thomson the leader in modern III-V compound semiconductor technology. Her laboratory was a dream for most of us, well before the common excellence of today in many places. But Razeghi became a technologist by choice. She was driven by the vision of the ultimate device backed by a deep understanding of the science and full of curiosity. This is what guided her. No wonder she became a very desired collaborator for top labs and personalities in the semiconductor world. She soon reached the peak of the Himalayas and could well have stopped there. But not for Mme Razeghi. After many years of success, she left friendly Europe for the next grand tour of her life, to the host of the most advanced materials science – the United States – interestingly, not to another industrial super-organization like Thomson, but to a university, where she could share her experiences and shape the next generations. Her energy and visions attracted money, and the money helped to create one of the most advanced university-based semiconductor labs in the world, visited and applauded by most Nobel laureates in the field.

So, dear readers, make sure that you learn from this book, but not only science and technology, which is presented with great clarity, skill, and care (there is even an appendix on how to work with dangerous chemicals in the MOCVD lab!). Maybe you will hear – just as I did – the whisper of the modestly hidden powerful message from Professor Razeghi: *the only thing to prevent you from performing miracles in the tournament with Nature is yourself*. To win and to have pleasure, learn first, then practice in the lab, and work with your notebook. If you work hard enough and still enjoy it, you may have the stuff for the ultimate destiny – real Himalayas – the discourse with nature: understand her laws and limitations, but also her immense and endless frontiers.

Thank you Manijeh for the guidance.

Professor in Physics, Institute of Physics
Polish Academy of Science, Warsaw, Poland
Fellow of the American Physical Society
Member of Academia Europaea

Jerzy M. Langer

Preface

Learning from Nature: Structure of Matter – Atoms

Nature is the best innovator and teacher. Scientists know for a while now that all matter consists of atoms. The atom is the smallest part of any material element. So when we look around us and observe the material world, we know that these natural colors we see are the light emitted by atoms. But atoms consist of nuclei surrounded by clouds of electron and the light particles they emit are what we call the quanta of light or photons. At the end of the last century, we learned from the great physicist James Clerk Maxwell that light, and its individual quanta, the photons are electromagnetic waves emanating from atomic emission or more generally from oscillating charges. Electrons undergo a transition from a higher to a lower orbit in an atom that emits light and conversely can also absorb light. The detail of this transition determines the energy or wavelength of the light. This includes the entire spectrum of light from gamma rays to UV to visible and to the invisible infrared (IR) rays down to the THz. Our eyes can see only a small part of the total photonic spectrum, from 300 to 700 nm in wavelength. So it is understandable that one of the first and primary aims of physicists was to try making instrumentation in order to see the rest of the spectrum as well, using artificial eyes. These electronic eyes are made by materials engineering. Indeed this has been achieved now to a great extent, and the progress is so important that artificial eyes covering a much larger range of photonic energies are being made and are constantly being improved. This progress was acquired by first developing a deep understanding of the workings of atoms. In fact one can say that the last century was the century of exploring the atom and mastering the science of materials. The next century will be the century of genes and biological cells.

Physicists have discovered that detecting and creating photons of different wavelengths require first a profound understanding of the atom, and this has been made possible by the science of quantum mechanics. The second step was to investigate a very special type of materials called the semiconductors. The science of semiconductors is central to all modern device physics, including the electronic chip and computer. Unlike in metals, electrical charges in semiconductors are not free to move under the action of a small electric field; they first have to be "excited,"

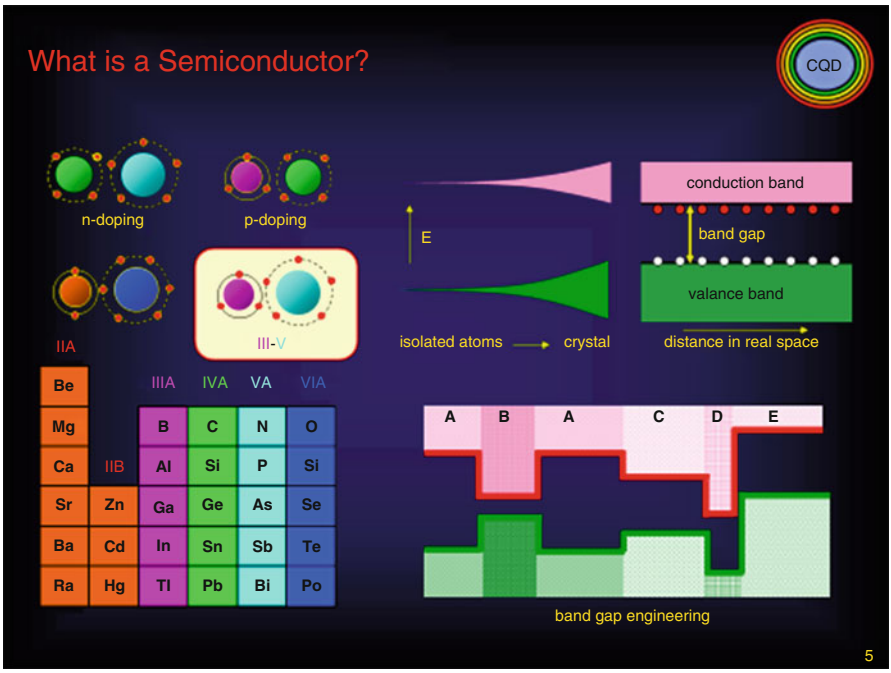


Fig. 1 Basic elements used to make pure and compound semiconductors; each has its own bandgap and, when combined into compounds, develops a new bandgap; layer-by-layer deposition generates new class of semiconductor superlattices with designer bandgaps (bottom right)

for example, by light or heat, to cross the energy gap formed by the bonding structure. This gap determines the sensitivity of the material to a particular wavelength and varies according to semiconductor type, and indeed the gap can now mostly be designed. To design and understand semiconductors, one has to realize that semiconductors, like other materials, consist of different types of atoms bonded together. Materials can be liquid, soft, or hard, and here we are in the first place talking about hard solids. The most useful and well-known semiconductors are in the category of silicon and germanium. What distinguishes them is that each atom has four valence electrons which combine their orbits to point to four different directions of space (tetrahedron) where they overlap and bond with four corresponding neighbor orbitals. These form semiconductors of the group IV-IV elements. Semiconductors can also be formed by combining group III and V elements such a GaAs or InAs (see Fig. 1). Here we have three and five valence electrons in the outer shells, respectively, and bonding comes about by first transferring an electron from element five to three, making it possible to form as in silicon, four tetrahedral bonds. There are many examples of III-V compounds, and they are extremely important to technology. Similarly, one can combine semiconductors by combining II-VI elements (CdTe, CdSe) where now two electrons are transferred from VI to II, making it again 4-4 bonds. A particularly inspiring and special atom is the atom of

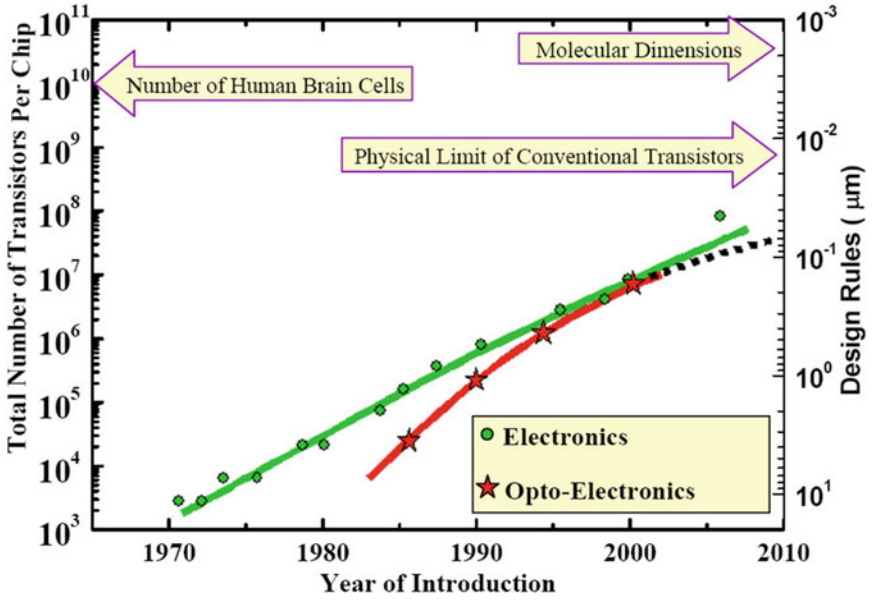


Fig. 2 Evolution of the total number of transistors per computer chip and their corresponding dimensions (in an inverted scale) as a function of year. For comparison, the number of human brain cells is shown on the left scale. In addition, the physical dimension limit for conventional transistors and the size of molecules are shown on the right scale

the element carbon. The four valence electrons of carbon can bond with one, two, or three neighbors and in this way form organic molecules, polymer chains, or two- or three-dimensional solids. A notable example is diamond which is bonded in three dimensions and is a high bandgap semiconductor with the highest thermal conductivity and hardness. The next example is the two-dimensional graphene (G). Graphene is causing a revolution in applied sciences. Carbon physics has already led to the awarding of three Nobel prizes, one for buckyballs (fullerenes) and the other two for graphene.

By now the reader should get a feeling of how exciting and useful solids and semiconductors and their applications are. But before we get into the details of how the solids work, what constitutes the important physics and engineering, and how we can develop the necessary sensory tools (see Fig. 2), let us revisit our own natural sensory systems and find out what challenge we are facing when we want to imitate or surpass nature.

Nature has stimulated human thought and invention before recorded time. Controlled fire, the wheel, and stone tools were all undoubtedly "invented" by humans, who drew inspiration from some natural phenomena in our prehistory, such as a wildfire created by a lightning strike, the rolling of round boulders down a steep hill, and perhaps wounds caused by the sharp rocks of a river bottom. There are examples during recorded times of other such ingenuity inspired by nature. Sir Isaac Newton

wrote that seeing an apple fall from a tree outside his window provoked his initial thoughts on the theory of gravitation. The Wright brothers and countless unsuccessful aviators before them were stimulated by the flight of birds. Similarly, we can look to nature to give us inspiration for new electronic devices.

Even a casual glance at the living world around us reveals the rich diversity and complexity of life on Earth. For instance, we can choose virtually any organism and demonstrate that it has the ability to sense and react to the surrounding world. Over millions of years of evolution, almost all types of life have developed some type of detection ability, seamlessly integrated into the other functions of the life form. More specifically, we can examine the basic human senses of hearing, smell, taste, touch, and sight to inspire us to understand more about the physical world.

Human hearing is based around the organ of Corti, which transduces pressure waves created within the fluids of the cochlea. The 20,000 micron-sized hair cells not only convert these waves into electrical impulses and transmit them to the brain via the auditory nerve but allow audio spectral differentiation depending on their position within the organ. Typical human frequency response ranges from 20 kHz to 30 Hz with sensitivity up to 130 decibels. Drawing from this natural example, today microphone manufacturers produce tiny transducers with dimensions of a few hundred microns.

The human sense of smell is based around approximately twelve million receptor cells in the nose. Each cell contains between 500 and 1000 receptor proteins that detect different scents and relay the information to the olfactory bulb and onto the brain. Today, researchers are developing "electronic noses" to mimic and improve upon the human olfactory system. Important applications include the detection of explosives as well as toxic chemicals and bio-warfare agents.

Gustatory receptors on the human tongue act as detectors for specific chemical molecules and are the basis for the sense of taste. Between 30,000 and 50,000 individual taste receptors make up the taste buds that cover the tongue and are capable of sensing bitter, sour, sweet, salty, and monosodium glutamate (MSG)-based foods. "Artificial tongues" are being developed to similarly classify flavors and also to perform specialized chemical analysis of a variety of substances. Aside from the obvious commercial applications (such as active sampling of foods and beverages in production), these devices may act in conjunction with "electronic noses" to detect various chemical agents for security purposes.

The sense of touch in humans allows several detection mechanisms, including specific receptors for heat, cold, pain, and pressure. These receptors are located in the dermis and epidermis layers of the skin and include specialized neurons that transmit electric impulses to the brain. Today, microswitches have been developed to detect very small forces at the end of their arms much like the whiskers of a cat. Thermocouples have been developed for sensitive temperature detection, and load cells are used for quantitative pressure sensing.

The sense of sight is perhaps the most notable form of human ability. Micron-sized rods and cones containing photosensitive pigments are located in the back of the eye. When light within the visible spectrum strikes these cells, nerves are fired and the impulses are transmitted through the optic nerve to the brain, with electrical

signals of only 100 mV between intracellular membranes. With the proper time to adapt to dark conditions, the human eye is capable of sensing at extremely low light levels (virtually down to single-photon sensitivity). However, our vision is limited to a spectral band of wavelengths between about 400 and 750 nanometers. In order to extend our sensing capabilities into the infrared and ultraviolet, much research has gone into exploring various material systems and methods to detect these wavelengths.

In order to improve and stretch the limits of innate human capabilities, researchers have mimicked nature with the development of quantum sensing techniques. Using these electronic noses, tongues, pressure sensors, and "eyes," scientists not only achieve a better understanding of nature and the world around them but also can improve the quality of life for humans. People directly benefit in a number of different ways from these advances ranging from restoration of sight, reduction in terrorist threats, and enhanced efficiency and speed of industrial processes.

Beyond human sensing capabilities, we can also look to the brain as an example of a computing and processing system. It is responsible for the management of the many sensory inputs as well as the interpretation of these data. Today's computers do a good job of processing numbers and are becoming indispensable in our daily lives, but they still do not have the powerful capabilities of the human brain. For example, state-of-the-art low power computer processors consume more power than a human brain while having orders of magnitude fewer transistors than the number of brain cells in a human brain (Fig. 1). Forecasts show that the current microelectronics technology is not expected to reach similar levels because of its physical limitations (Fig. 2).

By imitating nature, scientists have already developed a growing array of electronic sensors and computing systems. It is obvious that we must continue to take cues from the world around us to identify the proper methods to enhance human knowledge and capability. However, future advances in this direction will have to reach closer to the structure of atoms, by engineering *nanoscale electronics* (Fig. 3).

Thanks to nanoelectronics, it will not be unforeseeable in the near future to *create* artificial atoms, molecules, and integrated multifunctional nanoscale systems. For example, as illustrated in Fig. 4, the structure of an atom can be likened to that of a so-called quantum dot or Q-dot where the three-dimensional potential well of the quantum dot replaces the nucleus of an atom. An artificial molecule can then be made from artificial atoms. Such artificial molecules will have the potential to revolutionize the performance of optoelectronics and electronics by achieving, for example, orders of magnitude higher speed processors and denser memories. With these artificial atoms/molecules as building blocks, artificial active structures such as nanosensors, nanomachines, and smart materials will be made possible.

At the foundation of this endeavor is solid state engineering, which is a fundamental discipline that encompasses physics, chemistry, electrical engineering, materials science, and mechanical engineering. Because it provides the means to understand matter and to design and control its properties, solid state engineering is key to comprehend Natural Science (Fig. 5).

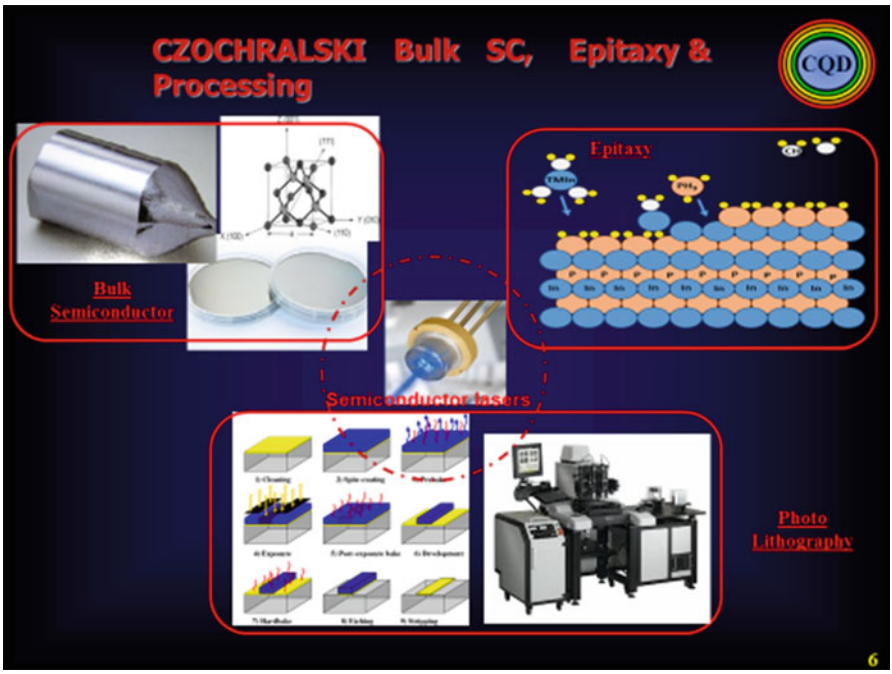


Fig. 3 The various ways a semiconductor is made as bulk (top left), with atomic beam deposition, and the way it is patterned and processed for device application using photolithography with laser beams to delimit regions

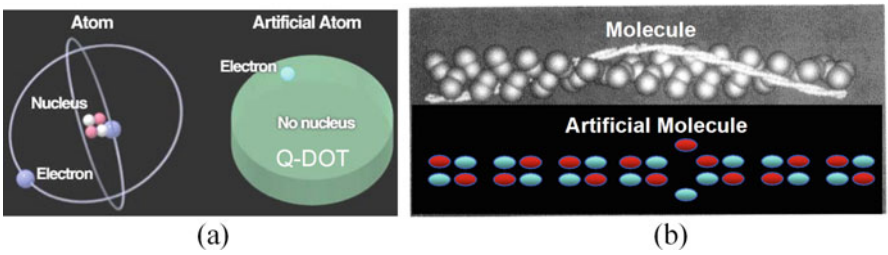


Fig. 4 Schematic comparisons: (**a**) between a real atom and an artificial atom in the form of a quantum dot and (**b**) between a real molecule and an artificial molecule

The twentieth century has witnessed the phenomenal rise of Natural Science and Technology into all aspects of human life. Three major sciences have emerged and marked that century, as shown in Fig. 3: Physical Science which has strived to understand the structure of atoms through quantum mechanics, Life Science which has attempted to understand the structure of cells and the mechanisms of life through biology and genetics, and Information Science which has symbiotically developed the communicative and computational means to advance Natural Science.

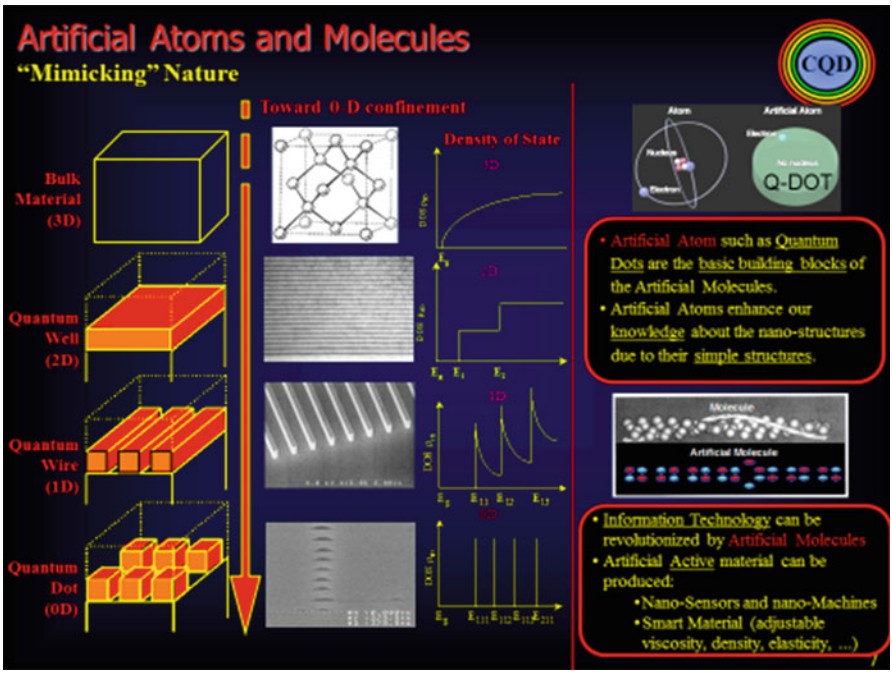


Fig. 5 The electronic structure and thus properties of materials, such as the density of available energy levels D(E), for example, changes with confinement size and dimensionality, and can be controlled by the great progress made using atom-by-atom deposition technologies such as MBE and MOCVD (molecular beam epitaxy and metalorganic chemical vapor deposition) areas in which the present author is a world leader

The scientific and technological accomplishments of earlier centuries represent the first stage in the development of Natural Science and Technology, that of understanding (Figs. 6 and 7). As the twenty-first century rolls in, we are entering the creation stage where promising opportunities lie ahead for creative minds to enhance the quality of human life through the advancement of science and technology.

Hopefully, by giving a rapid insight into the past and opening the doors to the future of solid state engineering, this course will be able to provide some of the basis necessary for this endeavor, inspire the creativity of the reader, and lead them to further explorative study.

Since 1992 when I joined Northwestern University as a faculty member and started to teach, I have established the Solid State Engineering (SSE) research group in the Electrical Engineering and Computer Science Department and subsequently created a series of related undergraduate and graduate courses. In the creative process for these courses, I studied similar programs in many other institutions such as Stanford University, the Massachusetts Institute of Technology, the University of Illinois at Urbana-Champaign, the California Institute of Technology, and the

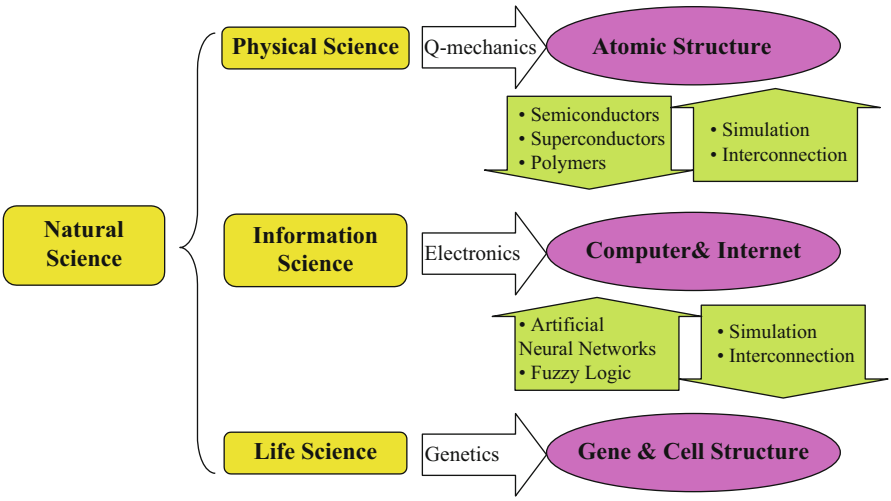


Fig. 6 Three branches of Natural Science and Technology have impacted all aspects of human life in the twentieth century: Physical, Information, and Life Sciences ( ). For each one, a key scientific discipline or technology has been developed: quantum mechanics, electronics, and genetics ( ). These have allowed to both better understand the building blocks of nature (structures of atoms, genes, and cells) and develop the tools without which these scientific advances would not have been possible (computer and Internet) ( ) in a synergetic manner ( )

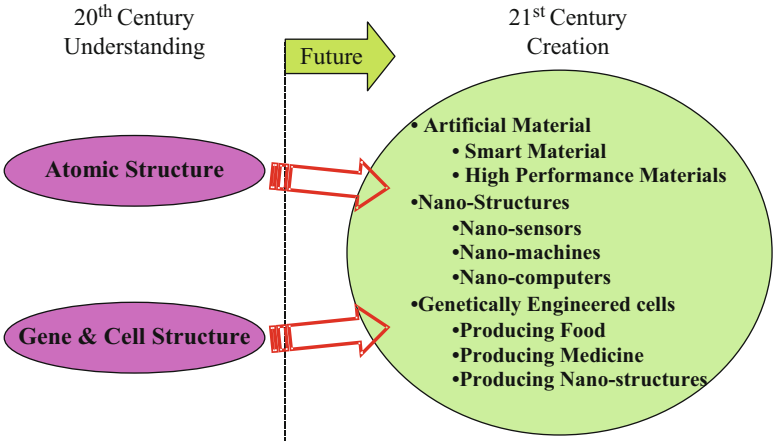


Fig. 7 The scientific and technological advances of the twentieth century can be regarded as the understanding stage in the development of Natural Science and Technology. The twenty-first century will be the creation stage in which novel opportunities will be discovered and carried out

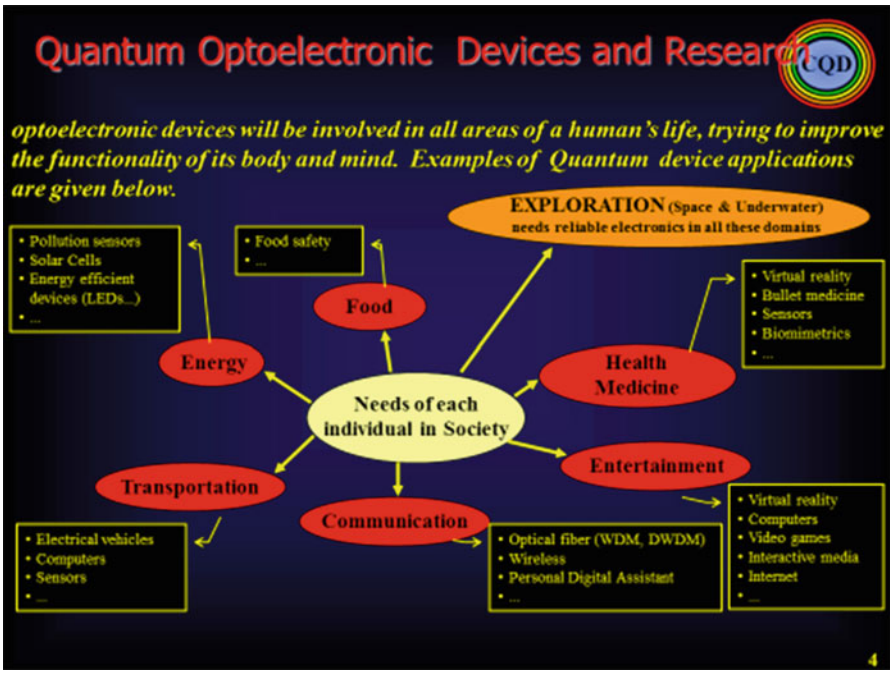


Fig. 8 The slide is self-explanatory: the many areas of life where solid state engineering has a direct impact

University of Michigan. I reviewed numerous textbooks and reference texts in order to put together the teaching material students needed to learn nanotechnology and semiconductor science and technology from the basics up to modern applications. But I soon found it difficult to find a textbook which combined all the necessary material in the same volume, and this prompted me to write the first edition of a textbook on the *Fundamentals of Solid State Engineering* (Fig. 8).

The book was primarily aimed at the undergraduate level, but graduate students and researchers in the field will also find useful material in it. After studying it, a student will be well versed in a variety of fundamental scientific concepts essential to solid state engineering, in addition to the latest technological advances and modern applications in this area, and will be well prepared to meet more advanced courses in this field.

In this fourth edition, I have taken into account the feedback and comments from students who took the courses associated with this text and from numerous colleagues in the field. The fourth edition is an updated, more complete text that covers an increased number of solid state engineering concepts and goes in depth in several of them. The chapters also include redesigned and larger problem sets.

This fourth edition is structured in two volumes. The first focuses on the basic physics concepts which are at the heart of solid state matter in general and semiconductors in particular. The text starts by providing an understanding of the

structure of atoms and electrons and the structure of matter (Chap. 1); a new chapter is devoted to the element carbon and its allotropes such as graphene, carbon nanotubes, and fullerenes (Chap. 2) and then the real and reciprocal crystal lattices (Chap. 3). An introduction to the basic concepts in quantum mechanics (Chap. 4) and to the modeling of electrons and energy band structures in crystals is then given in Chap. 5. Chapter 4 was extended in the fourth edition to include the Heisenberg equation of motion, the hydrogen atom, and the harmonic oscillator and quite a bit more. The new material now gives the student a reasonably complete description of the quantum mechanics tools that he needs. In Chap. 6 the attention is focused on the thermal and vibrational properties of crystals. The reader is introduced to the concept of phonons to describe vibrations of atoms in crystals (Chap. 6), and then later in the same chapter, he learns how to calculate the thermal properties of crystals. The equilibrium and non-equilibrium electrical properties of semiconductors are reviewed in Chaps. 7 and 8. First the statistics (Chap. 7) and then, later in Chap. 8, the transport description are developed. This now includes the Boltzmann equation approach. The problem of the generation and recombination properties of charge carriers in semiconductors is also considered in detail in Chap. 8. With these concepts one can now proceed to model semiconductor *p-n* and semiconductor-metal junctions (Chap. 9) which constitute the building blocks of modern electronics. The optical properties of semiconductors are described in Chap. 10. Solar, thermal, and photothermal harvesting of energy have been added in this new edition as Chaps. 11, 12, and 13. Screening and electron-electron interactions, on an elementary level, are the subjects of the new Chap. 14. This is followed by a discussion of semiconductor heterostructures and low-dimensional quantum structures including quantum wells and superlattices, wires, and dots in Chap. 15. The new Chap. 16 contains an introduction to the physics of quantum transport. A brief description of the coupling between electrons and lattice vibrations (electron-phonon interactions) then follows in the second part of Chap. 16. In the new and old chapters, the derivation of the mathematical relations has been spelled out in some detail, so that the reader can understand the limits of applicability of these expressions and adapt them to his or her particular needs. The final three chapters of the book focus on the growth and characterization of real semiconductor crystals. Chapter 17 introduces modern epitaxial and bulk semiconductor crystal growth techniques. This is followed by a discussion of semiconductor characterization techniques and defects in Chaps. 18 and 19.

In each chapter, a section "References" lists the bibliographic sources which have been referenced in the text. The interested reader is encouraged to read them in addition to those given in the section "Further reading."

Evanston, IL, USA Manijeh Razeghi

Contents

List of Symbols

a_0	Bohr radius
Å	Angstrom
α	Absorption coefficient
α_L	Thermal expansion coefficient
$\vec{B}$	Magnetic induction or magnetic flux density
c	Velocity of light in a vacuum
cal	Calorie
C, C_v, C_p	Heat capacity or specific heat, at constant volume, at constant pressure
d	Density
d	Distance, thickness or diameter
D, $\vec{D}$	Electric displacement
D, D_n, D_p	Diffusion coefficient or diffusivity, for electrons, for holes
Δn, Δp	Excess electron, hole concentration
$\vec{E}$	Electric field strength
E, E_n	Energy
E_C	Energy at the bottom of the conduction band
E_F	Fermi energy
E_{Fn}	Quasi-Fermi energy for electrons
E_{Fp}	Quasi-Fermi energy for holes
E_g	Bandgap energy
E_V	Energy at the top of the valence band
E_Y	Young's modulus
ε_0	Permittivity in vacuum
ε	Permittivity
ε_r	Dielectric constant
F, $\vec{F}$	Force
f	Frequency
f_e	Fermi-Dirac distribution for electrons
f_h	Fermi-Dirac distribution for holes
Φ_{ph}	Photon flux

Φ_B	Schottky potential barrier height
Φ_m, Φ_s	Work function of a metal, semiconductor
g	Gravitational constant
g	Density of states
G	Gibbs free energy
G, g	Gain
Γ	Optical confinement factor
H	Enthalpy
$\vec{H}$	Magnetic field strength
h	Planck's constant
$\hbar$	Reduced Planck's constant, pronounced "h bar", ($=h/2\pi$)
η	Quantum efficiency
η	Viscosity
i	$\sqrt{-1}$
i, I	Current
J, $\vec{J}$	Current density, current density vector
J^{diff}, $\vec{J}^{\text{diff}}$	Diffusion current density
J^{drift}, $\vec{J}^{\text{drift}}$	Drift current density
J_T	Thermal current
κ	Thermal conductivity coefficient
κ	Damping factor (imaginary part of the complex refractive index $\bar{N}$)
$\vec{K}$	Reciprocal lattice vector
k, $\vec{k}$	Wavenumber ($=2\pi/\lambda = 2\pi\nu/c$), wavenumber vector or wavevector
k_b	Boltzmann constant
k_D	Debye wavenumber
L_n, L_p	Diffusion length for electrons, holes
λ	Wavelength
Λ	Mean free path of a particle
m, M	Mass of a particle
m_0	Electron rest mass
m^*, m_e	Electron effective mass
m_h, m_{hh}, m_{lh}	Effective mass of holes, of heavy holes, of light holes
m_r^*	Reduced effective mass
M_V	Solid density (ratio of mass to volume)
μ	Permeability
μ_e	Electron mobility
μ_h	Hole mobility
n	Particle concentration
n	Electron concentration or electron density in the conduction band
n	Ideality factor in semiconductor junctions
$\bar{n}$	Refractive index (real part of the complex refractive index $\bar{N}$)

$\bar{N}$	Complex refractive index
N_{A}	Acceptor concentration
N_{c}	Effective conduction band density of states
N_{D}	Donor concentration
N_{v}	Effective valence band density of states
υ	Frequency
$\mathrm{N_A}$	Avogadro number
p	Hole concentration or hole density in the valence band
$p, \vec{p}$	Momentum
P	Power
Ψ	Wavefunction
q	Elementary charge
Q	Total electrical charge or total electrical charge concentration
ρ	Electrical resistivity
$\vec{r}$	Position vector
$\vec{R}$	Direct lattice vector
R	Resistance
R	Reflectivity
Ra	Rayleigh number
Re	Reynolds number
R_0	Differential resistance at $V = 0$ bias
R_{i}	Current responsivity
R_{v}	Voltage responsivity
Ry	Rydberg constant
S	Entropy
σ	Electrical conductivity
τ	Carrier lifetime
U	Potential energy
V	Voltage
$v, \vec{v}$	Particle velocity
v_{g}	Group velocity
ω	Angular frequency ($= 2\pi\upsilon$)
$\vec{x}, \vec{y}, \vec{z}$	Unit vectors (Cartesian coordinates)

Electronic Structure of Atoms

1

1.1 Introduction

In this chapter the electronic structure of single atoms will be discussed. A few quantum concepts will be introduced, as they are necessary for the understanding of many aspects in solid-state physics and device applications.

In Chap. 1, we saw that matter was composed of atoms in the periodic table shown in Fig. 1.2. Until 1911, atoms were considered the simplest constituents of matter. In 1911, it was discovered that atoms had a structure of their own and Rutherford proposed the nuclear model of the atom in which almost all the mass of the atom is concentrated in a positively charged nucleus and a number of negatively charged electrons are spread around the nucleus. It was later found that the nucleus is itself made up of protons (positively charged) and neutrons (neutrally charged). The number of protons is the atomic number (Z) while the total number of protons and neutrons is the mass number of the element. Apart from the electrostatic repulsion between nuclei, all of the major interactions between atoms in normal chemical reactions (or in the structures of elemental and compound substances) involve electrons. It is therefore necessary to understand the electronic structure of atoms. The term electronic structure, (or configuration) when used with respect to an atom, refers to the number and the distribution of electrons about the central nucleus.

The following discussion traces the steps of the scientific community toward a description of the electronic structure of atoms. The reader should not be stopped by the new concepts that arise from this discussion, because they will become clearer after understanding the quantum mechanics presented in Chap. 4.

Much of the experimental work on the electronic structure of atoms done prior to 1913 involved measuring the frequencies of electromagnetic radiation (e.g., light) that are absorbed or emitted by atoms. It was discovered that atoms absorbed or emitted only certain, sharply defined frequencies of electromagnetic radiation. These frequencies were also found to be characteristic of each particular element in the periodic table. And the absorption or emission spectra, i.e., the ensemble of

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_1

frequencies, were more complex for heavier elements. Before being able to understand the electronic structures of atoms, it was natural to start studying the simplest atom of all, the hydrogen atom, which consists of one proton and one electron.

1.2 Spectroscopic Emission lines and Atomic Structure of Hydrogen

It was experimentally observed that the frequencies of light emission from atomic hydrogen could be classified into several series. Within each series, the frequencies become increasingly closely spaced, until they converge to a limiting value. Rydberg proposed a mathematical fit to the observed experimental frequencies, which was later confirmed theoretically:

$$\frac{\upsilon}{c} = \frac{1}{\lambda} = Ry\left(\frac{1}{n^2} - \frac{1}{(n')^2}\right) \tag{1.1a}$$

with $n = 1, 2, 3, 4,...$ and $n' = (n + 1), (n + 2), (n + 3),...$

In this expression, λ is the wavelength of the light (in units of distance, and typically cm in this expression), υ is the frequency of the light emitted, c ($c = 2.99792 \times 10^8$ m·s^{-1} $= 2.99792 \times 10^{10}$ cm·s^{-1}) is the velocity of light in vacuum, and Ry is the fit constant, called the Rydberg constant, and was calculated to be 109,678 cm^{-1}. n is an integer, corresponding to each of the series mentioned above. n' is also an integer, larger than or equal to $(n + 1)$, showing that the frequencies become more closely spaced as n' increases.

The energy of the electromagnetic radiation is related to its wavelength and frequency by the following relation:

$$E = \frac{hc}{\lambda} = h\upsilon \tag{1.1b}$$

where h ($h = 6.62617 \times 10^{-34}$ J·s) is Planck's constant. The SI (Système International, or International System) unit for energy is the Joule (J). However, in solid-state physics, it is common to use another unit: the electron volt (eV) which is equal to 1.60218×10^{-19} J. The reason for this new unit will become clear later in the text and reflects the importance of the electron in solid-state physics.

The expression in Eq. (1.1a) shows that the emission of light from the hydrogen atom occurs at specific discrete values of frequencies ν, depending on the values of integers n and n'. The Lyman series of spectral lines corresponds to $n = 1$ for which the convergence limit is 109,678 cm^{-1}. The Balmer series corresponds to $n = 2$, and the Paschen series to $n = 3$. These are illustrated in Fig. 1.1, where the energy of the light emitted from the atom of hydrogen is plotted as arrows.

Although the absorption and emission lines for most of the elements were known before the turn of the twentieth century, a suitable explanation was not available, even for the simplest case of the hydrogen atom. Prior to 1913, the explanation for

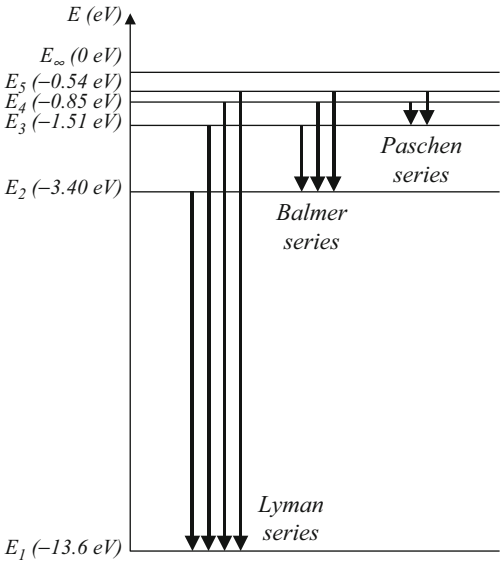


Fig. 1.1 Energies of the light emitted from the hydrogen atom (shown by arrows). The Lyman series corresponds to $n = 1$ in Eq. (1.1a), the Balmer series corresponds to $n = 2$, and the Paschen series corresponds to $n = 3$

this spectroscopic data was impossible because it contradicted the laws of nature known at the time. Indeed, very well-established electrodynamics could not explain two basic facts: that atoms could exist at all and that discrete frequencies of light were emitted and absorbed by atoms. For example, it was known that an accelerating charged particle had to emit electromagnetic radiation. Therefore, in the nuclear model of an atom, an electron moving around the nucleus has acceleration and thus has to emit light, lose energy, and fall down to the nucleus. This meant that the stability of elements in the periodic table, which is obvious to us, contradicted classical electrodynamics. A new approach had to be followed in order to resolve this contradiction, which resulted in a new theory, known as quantum mechanics. Quantum mechanics could also explain the spectroscopic data mentioned above and adequately describe experiments in modern physics that involve electrons and atoms and ultimately solid-state device physics.

Niels Bohr first explained the atomic absorption and emission spectra in 1913. His reasoning was based on the following assumptions, which cannot be justified within classical electrodynamics:

1. Stable orbits (states with energy E_n) exist for an electron in an atom. While in one of these orbits, an electron does not emit any electromagnetic radiation. An individual electron can only exist in one of these orbits at a time and thus has an energy E_n.
2. The transition of an electron from an atomic orbit of energy state E_n to that of energy state $E_{n'}$ corresponds to the emission ($E_n > E_{n'}$) or absorption ($E_n < E_{n'}$) of electromagnetic radiation with an energy $|E_n - E_{n'}|$ or frequency $\upsilon = \frac{|E_n - E_{n'}|}{h}$.

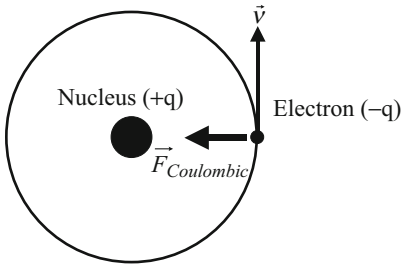


Fig. 1.2 Schematic diagram showing the electron orbit, the attractive coulombic force between the positively charged nucleus and the orbiting negatively charged electron, and the velocity of the electron which is always tangential to its circular orbit

With Sommerfeld, Bohr implemented these postulates into a simple theory. Assumption (1) of stable orbits meant that the values of angular momentum L and thus the electron orbit radius $\vec{r}$ were quantized, i.e., integer multiples of a constant. For the simple hydrogen atom with a circular electron orbit, the Bohr postulate (1) can be expressed mathematically in the following manner:

$$L_n = m\nu r_n = n\frac{h}{2\pi}, \quad n = 1, 2, \ldots \tag{1.2a}$$

where m is the mass of the electron, ν is the linear electron velocity, and n is an integer expressing the quantization and used to index the electron orbits. Since the orbit is circular, the electron experiences a centripetal acceleration ν^2/r_n. The coulombic force between the electron and nucleus provides this acceleration, as illustrated in Fig. 1.2.

Therefore, according to Newton's second law, equating Coulomb force with the mass times the centripetal acceleration, we can write:

$$\frac{q^2}{4\pi\varepsilon_0 r_n^2} = \left|\vec{F}_{\text{coulombic}}\right| = \frac{m\nu^2}{r_{\text{n}}} \tag{1.2b}$$

where ε_0 ($\varepsilon_0 = 8.85418 \times 10^{-12}$ F·m^{-1}) is the permittivity of free space and q ($q = 1.60218 \times 10^{-19}$ C) is the elementary charge.

Combining Eqs. (1.2a and 1.2b), one obtains the discrete radius of an electron orbit:

$$r_n = \frac{\varepsilon_0 n^2 h^2}{\pi m q^2} \tag{1.3}$$

The total electron energy E_n in the various orbits is the sum of the kinetic and (coulombic) potential energies of the electron in the particular orbit:

$$E_n = \frac{1}{2}\frac{q^2}{4\pi\varepsilon_0 r_n} - \frac{q^2}{4\pi\varepsilon_0 r_n} = -\frac{1}{8}\frac{q^2}{\pi\varepsilon_0 r_n} \tag{1.4}$$

With Eq. (1.3) we finally have:

$$E_{\text{n}} = \frac{-m\,q^4}{8\,(\varepsilon_0 n\,h)^2} = -\frac{13.6}{n^2} \text{ in units of electron} - \text{volts (eV)} \tag{1.5}$$

This theory thus provided an explanation for each series of spectroscopic lines in the emission spectrum from atomic hydrogen as shown in Fig. 1.1. An electron has the lowest (i.e., most negative) energy when it is in the orbit $n = 1$. The radius of this orbit can be calculated using Eq. (1.3) and is $a_0 = 0.52917$ Å. If an electron is excited to an orbit with higher energy ($n' \geq 2$) and returns to the ground state ($n = 1$), electromagnetic radiation with the frequency $c \times Ry\left[\left(\frac{1}{1^2}\right) - \left(\frac{1}{n'^2}\right)\right]$ is emitted, where c is the velocity of light in vacuum and Ry the Rydberg constant. In this case, the Lyman series of spectroscopic lines in Fig. 1.1 is observed. The other series arise when the electron drops from higher levels to the levels with $n = 2$ (Balmer series) and $n = 3$ (Paschen series), as shown in Fig. 1.1. Therefore, the Bohr-Sommerfeld theory could accurately interpret the observed, discrete absorption/emission frequencies in the hydrogen atom. Despite its success for the hydrogen atom, this theory still had to be improved for a number of reasons. One major reason was that it could not successfully interpret the spectroscopic data for atoms more complex than hydrogen. However, the results of Bohr's model can be extended to other structures similar to the hydrogen atom, called hydrogenoid systems. For example, the energy levels of several ionized atoms that have only a single electron (e.g., He^+ or Li^{++}) can be easily predicted by substituting the nuclear charge q of Bohr's model with Zq where Z is the atomic number.

The simple picture developed by Niels Bohr for electrons in atoms was among the first attempts to explain experimental data with assumptions based on the discrete (or quantum) nature of the electromagnetic field.

A typical example of the interaction between an electromagnetic field and matter is a blackbody, which is an ideal radiator of electromagnetic radiation. Using classical arguments, Rayleigh and Jeans tried to explain the observed blackbody spectral irradiance, which is the power radiated per unit area per unit wavelength, shown in Fig. 1.3. However, as can be seen in the figure, their theoretical predictions could only fit the data at longer wavelengths. In addition, their results also indicated that the total irradiated energy (integral of the irradiance over all the possible wavelengths) should be infinite, a fact that was in clear contradiction with experiment. In 1901, Max Planck provided a revolutionary explanation based on the hypothesis that the interaction between atoms and the electromagnetic field could only occur in discrete packets of energy, thus showing that the classical view that always allows a continuum of energies was incorrect. Based on these ideas, a more sophisticated and self-consistent theory was created in 1920 and is now called quantum mechanics (see Chap. 4 for more details).

1.3 Atomic Orbitals

Bohr's model solved the problem of the energy levels in the hydrogen atom but had several drawbacks: it could neither explain some of the other properties of hydrogen atoms nor correctly predict the energy levels of more complex atoms. In addition, a few years later, new experiments pointed out that particles could behave as waves,

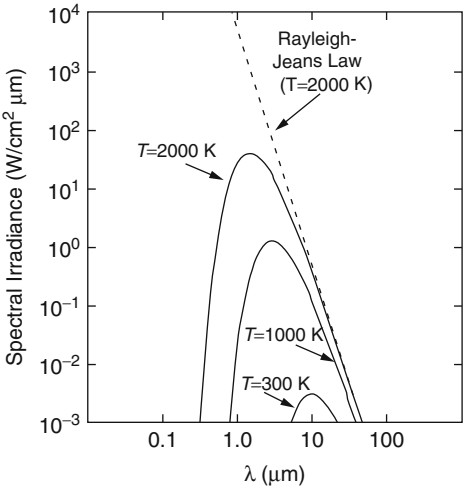


Fig. 1.3 Spectral irradiance of a blackbody at different temperatures. When the temperature is at or below room temperature, the radiation is mostly in the infrared spectral region, undetectable by the human eye. When the temperature is raised, the emission power increases and its peak shifts toward shorter wavelengths. One of the more successful interpretations, yet inaccurate because it was based on classical mechanics, was conducted by Rayleigh and Jeans but could only fit the experimental data at longer wavelengths

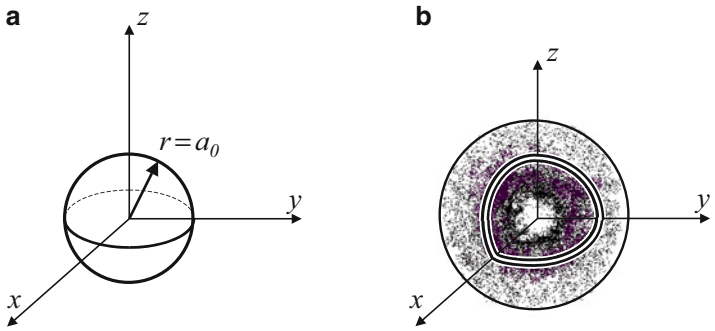


Fig. 1.4 (**a**) The precise spherical orbit of an electron in the first Bohr orbit, for which the radius is $a_0 = 0.52917$ Å, as calculated by Bohr's model. (**b**) The electron probability density pattern for the comparable atomic orbital using a quantum mechanical model. The darker areas indicate a higher probability of finding the electron at that location. The center cutout shows the interior of the orbital. The outer sphere delineates the region where the electron exists 90% of the time

and therefore their position could not be determined exactly. In Bohr's model, the radius of the first Bohr orbit in the hydrogen atom was calculated to be exactly $a_0 = 0.52917$ Å (Angstrom, abbreviated as Å, is equal to 10^{-10} m). This distance is a constant called the Bohr radius and is shown in Fig. 1.4a as a spherical surface with radius a_o.

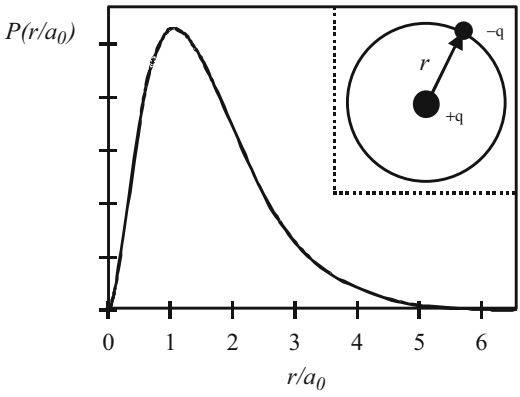


Fig. 1.5 The electron radial probability density function $P(r/a_0)$, which describes the probability of finding an electron in a spherical surface at a distance r from the nucleus in the hydrogen atom (for $n = 1$). This probability has a maximum value when the electron is at a distance equal to the Bohr radius: $r = a_0$

A new approach was clearly needed in order to describe matter on the atomic scale. This new approach was elaborated during the next decade and is now called quantum mechanics. In quantum mechanics an electron cannot be visualized as a point particle orbiting with a definite radius, but rather as a delocalized cloud with inhomogeneous probability density around a nucleus as illustrated in Fig. 1.4b. The probability density gives the probability of finding the electron at a particular point in space. In this picture, the Bohr radius can be interpreted as the radius a_0 of the spherical surface where the maximum in the electron probability distribution occurs or, in other words, the spherical orbit where the electron is most likely to be found. This can be further illustrated by Fig. 1.5 where the electron probability density function $P(r)$, which is the probability to find an electron at a distance r from the nuclei, is plotted as a function of r (for the lowest energy state of hydrogen atom $n = 1$). This function reaches its maximum at the value of Bohr's first orbit a_0.

We saw earlier that there were several stable orbits for an electron in the hydrogen atom which are distinguished by the energy given in Eq. (1.5). The orbit or energy is not enough to characterize the properties of an electron in an atom. The spatial shape and direction of the orbit are also important, as it is not always spherical, and so the term "orbital" is employed. Each orbital is assigned a unique set of quantum numbers, which completely specifies the orbital's properties. The orbital designation and its corresponding set of three quantum numbers n, l, and m_l are listed in Table 1.1 along with the electron spin quantum number m_s.

The principal quantum number n may take integral values from 1 to ∞, although values larger than 7 are spectroscopically and chemically unimportant. It is the value of this quantum number n that determines the size and energy of the principal orbitals. Orbitals with the same n are often called "shells."

For a given value of n, the angular momentum quantum number l may take integer values within [0, 1, 2, 3, …, $(n-1)$]. It is this quantum number that determines the shape of the orbital. A letter designation is used for each orbital shape: s for ($l = 0$), p for ($l = 1$), d for ($l = 2$), f for ($l = 3$), etc. followed alphabetically by the letter designations g, h, and so on.

Table 1.1 Quantum numbers and atomic orbital designations for electrons in the four lowest values of n. When n increases, the scheme continues to develop with the same basic rules

Orbital	n	l	m_l	m_s
1 s	1	0	0	–½, +½
2 s	2	0	0	–½, +½
2p	2	1	–1, 0, +1	–½, +½
3 s	3	0	0	–½, +½
3p	3	1	–1, 0, +1	–½, +½
3d	3	2	−2, −1, 0, +1, +2	–½, +½
4 s	4	0	0	–½, +½
4p	4	1	−1, 0, +1	–½, +½
4d	4	2	−2, −1, 0, +1, +2	–½, +½
4f	4	3	−3, −2, −1, 0, +1, +2, +3	–½, +½

Finally, for a given orbital shape (i.e., a given value of l), the magnetic quantum number m_l may take integral values from $-l$ to $+l$. This quantum number governs the orientation of the orbital. Once an electron is placed into one specific orbital, its values for the three quantum numbers n, l, and m_l are known.

A fourth quantum number is needed to uniquely identify an electron in an orbital, the spin quantum number. The spin quantum number is independent of the orbital quantum numbers and can only have two opposite values: $m_s = \pm\frac{1}{2}$ (in units of $\frac{h}{2\pi}$). Electrons that differ only in their spin value can only be distinguished in the presence of an external magnetic field.

1.4 Structures of Atoms with Many Electrons

In multi-electron atoms, the energy of an electron depends on the orbital principal quantum number n and the orbital momentum quantum number l, i.e., whether the electron is in an s, p, d, or f state. The different m_l quantum numbers for a fixed set of n and l are degenerate (they have the same energy). The electronic configurations of such atoms are built up from the ground state energy, filling the lowest energy orbitals first. Then, the filling of the orbitals occurs in a way such that no two electrons may have the same set of quantum numbers. This rule governing electron quantum numbers is called the Pauli exclusion principle. If two electrons occupy the same orbital, they must have opposite spins: $m_s = +½$ for one electron and $m_s = -½$ for the other electron. Because the spin quantum number m_s can take only these two values, an orbital with given *(n, l, m)* can be occupied by at most two electrons.

One more rule, called Hund's rule, governs the electron configuration in multi-electron atoms: for a given principal quantum number n, the lowest energy electron configuration has the greatest possible sum of spin values and greatest sum of orbital momentum values.

Example

Q Hund's rule says that the electrons occupy orbitals in such a way that, first, the total spin number ($\sum m_s$) is maximized and then the total orbital momentum is

maximized ($\sum l$). Determine the electronic configuration, including the spin, of the carbon atom, which has six electrons in its ground state.

A Carbon has six electrons and has the electronic configuration $1s^2 2s^2 2p^2$. The last two electrons in the p shell can have spins +½ or –½. To maximize the total spin number, both electrons must have their spin up, so that $\sum m_s = 1$, as shown below.

Incorrect Correct

Both the Pauli exclusion principle and Hund's rule govern the electron configurations of atoms in the periodic table in their unexcited state, which is also called the ground state. Other electronic configurations are possible when the atom is in an excited state as a result of an external force such as an electric field.

Examples of the ground state electron configurations in a number of elements are shown below. The sequence for $Z = 1$ to $Z = 18$ is built in a straightforward and logical manner, by filling the allowed s, p, d... orbitals successively (i.e., in this order). For $Z = 19$, the first deviation to this procedure occurs: the 4s orbitals are filled with electrons *before* the 3d orbitals. Elements in the periodic table with partially filled 3d orbitals are usually transition metals and the electrons in these 3d orbitals contribute to the magnetic properties of these elements. For example, the electronic configuration of the Ga element can be read as follows: two s-electrons in orbit 1, two s-electrons in orbit 2, six p-electrons in orbit 2, two s-electrons in orbit 3, six p-electrons in orbit 3, two s-electrons in orbit 4, ten d-electrons in orbit 3, and one p-electron in orbit 4.

$Z = 3\text{Li} \quad 1s^2 2s^1$
$Z = 4\text{Be} \quad 1s^2 2s^2$
$Z = 5\text{B} \quad 1s^2 2s^2 2p^1$
$Z = 6\text{C} \quad 1s^2 2s^2 2p^2$
$Z = 7\text{N} \quad 1s^2 2s^2 2p^3$
$Z = 8\text{O} \quad 1s^2 2s^2 2p^4$
$Z = 9\text{F} \quad 1s^2 2s^2 2p^5$
$Z = 10\text{Ne} \quad 1s^2 2s^2 2p^6$
... ...
$Z = 31\text{Ga} \quad 1s^2 2s^2 2p^6 3\text{s}^2 3\text{p}^6 4s^2 3d^{10} 4p^1$

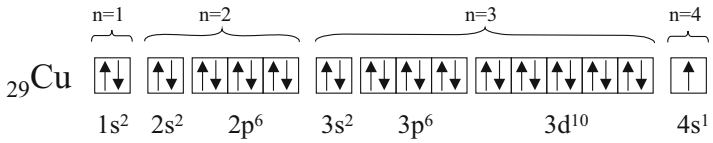
Example

Q Determine the electronic configuration for copper (element Cu, atomic number $Z = 29$ in the ground state).

A There are 29 electrons in copper in its ground state. It has an inner Ar shell, which has 18 electrons: [Ar] = $1s^2 2s^2 2p^6 3s^2 3p^6$. The remaining 11 electrons must be distributed inside the 3d and 4s orbitals. Suppose that the two

possible configurations are [Ar]$3d^9 4s^2$ and [Ar]$3d^{10} 4s^1$. According to Hund's rule, the lowest energy configuration, corresponding to the ground state, is such that it presents the greatest possible spin value and greatest orbital momentum. The two configurations above have the same spin but the second one has greater orbital momentum. Since the orbital quantum number for the *s* orbital is 0 and for *d* is 2, we can say that Cu exhibits the second electronic configuration:

[Ar]$3d^{10} 4s^1$ or $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^1$ which is illustrated below:



Quantum mechanics is able to predict the energy levels of the hydrogen atom, but the calculations become too complex for atoms with two or more electrons. In multi-electron atoms, the electric field experienced by the outer shell electrons does not correspond to the electric field from the entire positive nuclear charge because other electrons in inner shells screen this electric field from the nucleus. This is why outer shell electrons do not experience a full nuclear charge *Z* (the atomic number), but rather an effective charge *Z** which is lower than *Z*. Values of the effective nuclear charge *Z** for the first ten elements are listed in Table 1.2. Therefore, the energy levels of these outer shell electrons can be estimated using the results from the hydrogen atom and substituting the full nuclear charge *Zq* with *Z*q*.

Let us consider an example of electronic configuration in the multi-electron atom of Si. As shown in Fig. 1.6, 10 of the 14 Si-atom electrons (2 in the 1*s* orbital, 2 in the 2*s* orbital, and 6 in the 2*p* orbital) occupy very low energy levels and are tightly bound to the nucleus of the atom. The binding is so strong that these ten electrons remain essentially unperturbed during most chemical reactions or atom-atom

Table 1.2 The full nuclear charge Z and effective nuclear charge Z^* for the first ten elements

Element	*Z*	Z^*
H	1	1.00
He	2	1.65
Li	3	1.30
Be	4	1.95
B	5	2.60
C	6	3.25
N	7	3.90
O	8	4.55
F	9	5.20
Ne	10	5.85

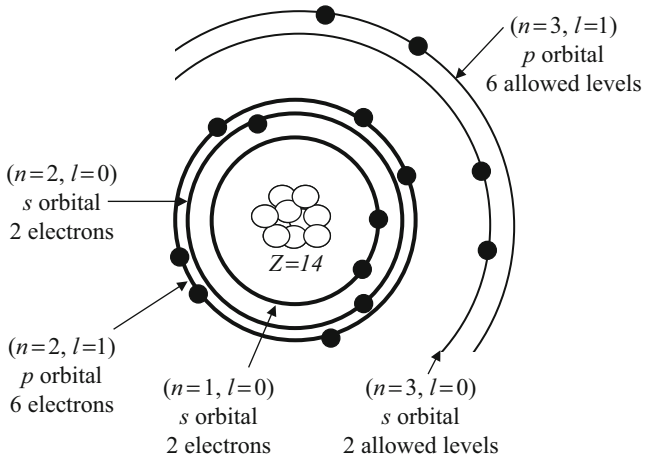
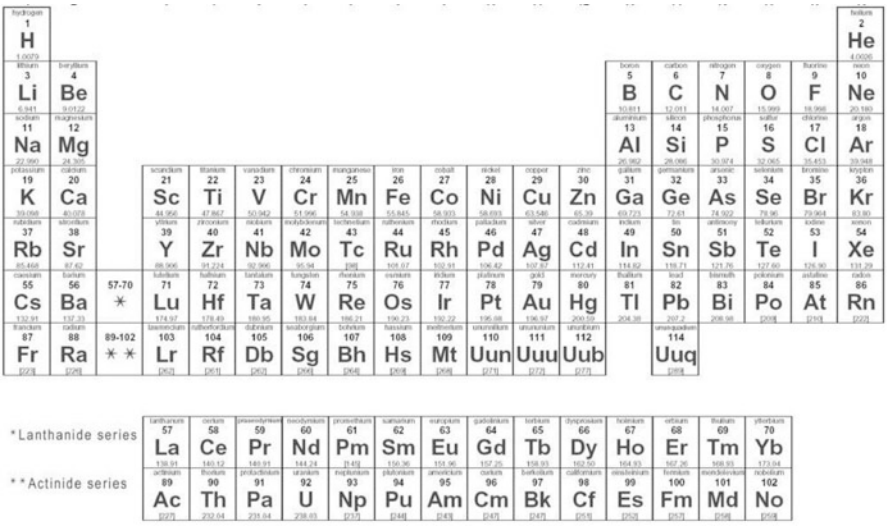


Fig. 1.6 Electron configuration for electrons in a Si atom. The ten electrons in the core orbitals, 1s ($n = 1$), 2s ($n = 2, l = 0$), and 2p ($n = 2, l = 1$) are tightly bound to the nucleus. The remaining four electrons in the 3s ($n = 3$, $l = 0$) and 3p ($n = 3$, $l = 1$) orbitals are weakly bound

interactions. The combination of the ten-electron-plus-nucleus is often being referred to as the "core" of the atom. On the other hand, the remaining four Si-atom electrons are rather weakly bound and are called the valence electrons because of their strong participation in chemical reactions. Valence electrons are those in the outermost occupied atomic orbital. As emphasized in Fig. 1.6, the four valence electrons occupy four of the eight allowed states belonging to the 3s and 3p orbitals.

The electronic configuration in the 32-electron Ge-atom (germanium being the next elemental semiconductor in column IV of the periodic table) is essentially identical to the Si-atom configuration except that the Ge-core contains 28 electrons.



1.5 Bonds in Solids

1.5.1 General Principles

When two atoms are brought very close together, the valence electrons interact with each other and with the neighbor's positively charged nucleus. As a result, a bond between the two atoms forms, producing, for example, a molecule. The formation of a stable bond means that the energy of the system of two atoms kept together must be less than that of the system of two atoms kept apart, so that the formation of the pair or the molecule is energetically favorable. Let us view the formation of a bond in more detail.

As the two atoms approach each other, they are under attractive and repulsive forces from each other as a result of mutual electrostatic interactions. At most distances, the attractive force dominates over the repulsive force. However, when the atoms are so close that the individual electron shells overlap, there is very strong proton-to-proton shell repulsion, called core repulsion, that dominates. Figure 1.7 shows the interatomic interaction energy as a function of the distance between atoms r. The system has zero energy when the atoms are infinitely far apart. A negative value corresponds to an attractive interaction, while a positive value stands for a repulsive one. The resulting interaction is the sum of the two and has a minimum at an equilibrium distance, which is reached when the attractive force balances the repulsive force. This equilibrium distance is called the equilibrium separation and is

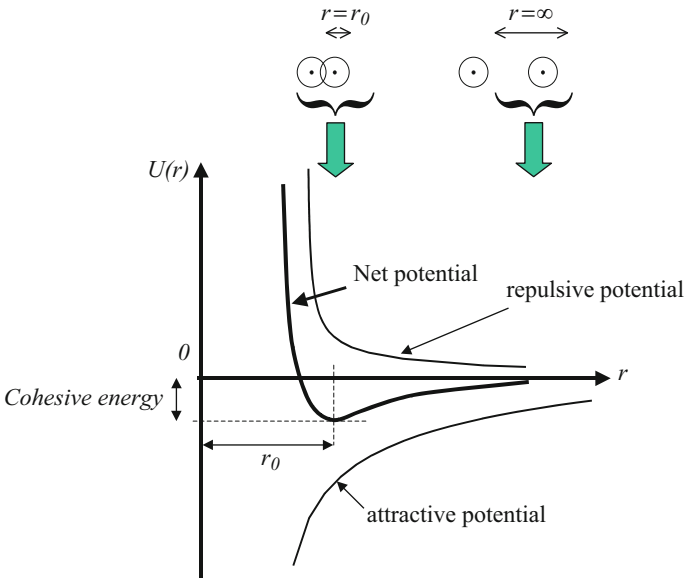


Fig. 1.7 Potential energy versus interatomic separation r. The net potential is the sum of repulsive and attractive components. The minimum of the net potential corresponds to the equilibrium distance r_0 between the two atoms

effectively the bond length. The energy required to separate the two atoms represents the cohesive energy or bond formation energy or simply bond energy (also shown in Fig. 1.7).

Similar arguments also apply to bonding between many more atoms, such as the billions of atoms found in a typical macroscopic solid. Even in the presence of many interacting atoms in a solid, we can still identify a general potential energy curve $U(r)$ per atom similar to the one shown in Fig. 1.7. Although the actual details will change from material to material, the general concepts of bond energy U_0 per atom and equilibrium interatomic separation will still be valid. These characteristics determine many properties of solids such as the thermal expansion coefficient and elastic modulus.

Example

Q For a face-centered cubic lattice, such as in an inert gas turned solid at low temperature, the potential energy can be expressed as:

$$U = N\left[12.13\left(\frac{\sigma}{r}\right)^{12} - 14.45\left(\frac{\sigma}{r}\right)^{6}\right],$$

where r is the distance between nearest neighbors and σ is a constant of the crystal. Determine the lattice constant a of the lattice in terms of σ.

A The equilibrium distance r is given by the minimum of the potential energy, which can be calculated by taking the derivative of the function U with respect to r and setting it equal to zero:

$$\frac{dU}{dr} = N\left[-145.56\frac{\sigma^{12}}{r^{13}} + 86.7\frac{\sigma^{6}}{r^{7}}\right] = 0$$

which yields $r = 1.09\sigma$. Since we are considering a face-centered cubic lattice, the nearest neighbor distance is such that $r = \frac{\sqrt{2}}{2}a$. Therefore, the lattice constant is $a = 1.54\sigma$.

1.5.2 Ionic Bonds

When one atom completely loses a valence electron so that the outer shell of a neighboring atom becomes completely filled, a bond is formed which is called ionic bond. The coulombic attraction between the now ionized atoms causes the ionic bonding. NaCl salt is a classic (and familiar) example of a solid in which the atoms are held together by ionic bonding. Ionic bonding is frequently found in materials that normally have a metal and a nonmetal as the constituent elements. For example, Fig. 1.8 illustrates the NaCl structure with valence electrons shifted from Na atoms to Cl atoms forming negative Cl^- ions and positive Na^+ ions. The physical structure of the NaCl crystal is shown in Fig. 1.9.

Ionic bonds generally have bond energies on the order of a few eV. The energy required to take solid NaCl apart into individual Na and Cl atoms is the cohesive energy, which is 3.15 eV per atom. The attractive part of Fig. 1.7 can be estimated from the sum of the coulombic potential energies between the ions (see Problem 11).

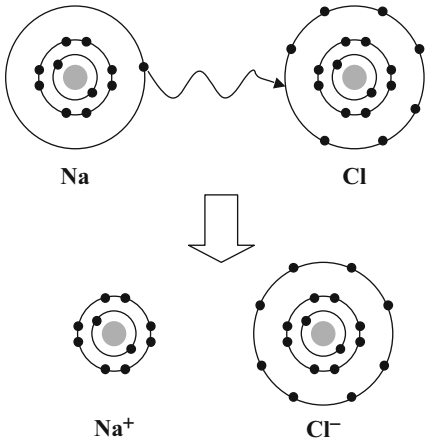


Fig. 1.8 Schematic illustration of the formation of an ionic bond in NaCl, showing the electron transfer between the two elements and their final electronic configurations

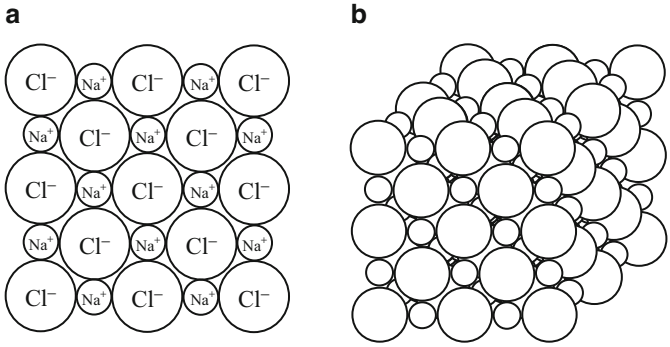


Fig. 1.9 (**a**) A schematic illustration of a cross-section from solid NaCl. Solid NaCl is made from Cl^- and Na^+ ions arranged alternatively, so that the oppositely charged ions are closest to each other and attract each other. There are also repulsive forces between the like-ions. In equilibrium, the net force acting on any ion is zero. (**b**) 3D illustration of solid NaCl

Example

Q Calculate the total coulombic potential energy of a Cs^+ ion in a CsCl crystal by only considering the nearest neighbors of Cs^+.

A In the cubic unit cell shown in Fig. 1.9, one can see that one Cs^+ ion (at the center of the cube) has eight nearest Cl^- neighbors (at the corners of the cube). Since the lattice constant for CsCl is $a = 4.11$ Å, the distance between a Cs^+ and one of its Cl^- neighbors is $r_{nn} = \frac{\sqrt{3}}{2}a = 3.56$ Å. The coulombic potential energy is thus $E = -8\frac{q^2}{4\pi\epsilon_0 r_{nn}} = -32.36$ eV.

Many other solids consisting of metal-nonmetal elements also have ionic bonds. They are called ionic crystals and, by virtue of their ionic bonding characteristics, share many similar physical properties. For example, LiF, MgO (magnesia), CsCl,

and ZnS are all ionic crystals; they are strong, brittle materials with high melting temperatures compared to metals. Most are soluble in polar liquids such as water. Since all the electrons are within the rigidly positioned ions, there are no free electrons to move around in contrast to metals. Therefore, ionic solids are typically electrical insulators. Compared to metals and covalently bonded solids, ionically bonded solids also have poor thermal conductivity.

1.5.3 Covalent Bonds

Two atoms can form a bond with each other by sharing some or all of their valence electrons and thereby reducing the overall energy. This is in contrast with an ionic bond because the electrons are shared rather than completely transferred. This concept is purely quantum mechanical and has no simple classical analogue. Nevertheless, it still results in the same basic principles as those shown in Fig. 1.7, i.e., there is a minimum in the total potential energy at the equilibrium position $r = r_0$.

Covalent bonds are very strong in solids. Figure 1.10 shows the formation of a covalent bond between atoms in crystalline Si, which has the diamond structure with eight atoms per cubic unit cell. Each Si shares its four valence electrons with its neighbors as shown in Fig. 1.10. There is an electron cloud in the region between atoms equivalent to two electrons with opposite spins.

In the structure of diamond, a C atom also shares electrons with other C atoms. This leads to a three-dimensional network of a covalently bonded structure as shown in Fig. 1.11. The coordination number (CN) is the number of nearest neighbors for a given atom in the solid. As it is seen in Fig. 1.11, the coordination number for a carbon atom in the diamond crystal structure is four, as discussed in Chap. 2.

In the tetrahedral systems such as C, Si, or Ge, for example, the covalent bonds undergo a very interesting process called hybridization. What happens is that the atom first promotes one of outer *s*-electrons (e.g., 2*s* shell in C and 3*s* shell in Si) into the doubly occupied *p*-shell. This costs energy, but this energy is more than

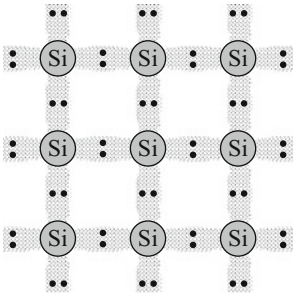


Fig. 1.10 Schematic of covalent bonds in Si. Each Si atom contributes one of its four outer shell electrons with each neighboring Si atom. This creates a pair of shared electrons between two Si atoms, which constitutes the covalent bond. Because the two atoms are identical, the electrons have the highest probability of being located at equal distances between the two atoms, as illustrated here

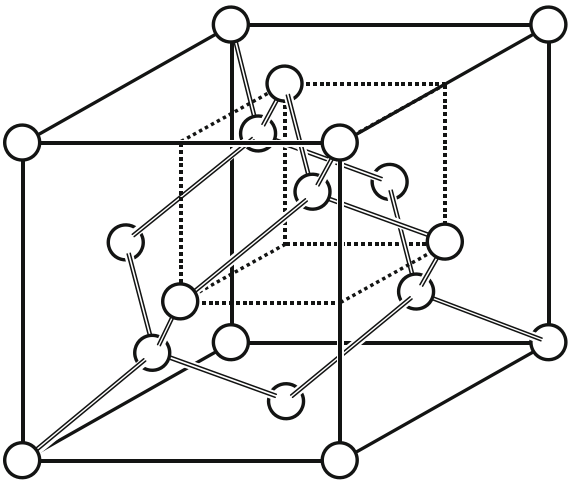


Fig. 1.11 The diamond crystal with covalent bonds. The diamond crystal is most often represented using a cubic unit cell, as shown here. Each atom in the structure is covalently bonded to four neighboring atoms

recovered because now the system can use the $2p_x$, $2p_y$, $2p_z$ orbitals in C, for example, to combine with the one left over in "s" to form four directed bonds:

$$\frac{1}{2}(2s + 2p_x + 2p_y + 2p_z)$$

$$\frac{1}{2}(2s + 2p_x - 2p_y - 2p_z)$$

$$\frac{1}{2}(2s - 2p_x + 2p_y - 2p_z)$$

$$\frac{1}{2}(2s - 2p_x - 2p_y + 2p_z)$$

pointing toward the four other atoms, where the same process has taken place, each atom providing a bond partner which is pointing in the opposite direction and giving maximum overlap.

Due to the strong Coulomb attraction between the shared electrons and the positive nuclei, the covalent bond energy is the strongest of all bond types, leading to very high melting temperatures and very hard solids: diamond is one of the hardest known materials. Covalently bonded solids are also insoluble in nearly all solvents. The directional nature and strength of the covalent bond also makes these materials nonductile (or nonmalleable). Under a strong force, they exhibit brittle fracture.

1.5.4 Mixed Bonds

In many solids, the bonding between atoms is generally not just of one certain type but rather is a mixture of bond types. We know that bonding in silicon is totally covalent, because the shared electrons in the bonds are equally attracted by the

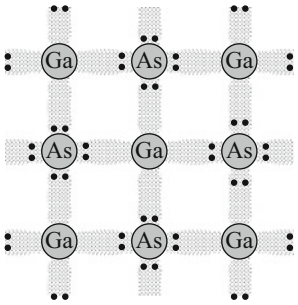


Fig. 1.12 Polar bonds in a III–V intermetallic compound. Similar to the case of Si in Fig. 1.10, a covalent bond is formed by the sharing of an electron from a Ga atom and one from a neighboring As atom. However, because a Ga atom has only three electrons in its outer shell, while an As atom has five, one of the four covalent bonds is formed by the As atom contributing two electrons, while the Ga atom contributes none. In addition, because the atoms involved are not the same, the electrons in the bonds are more attracted toward the atom with largest nucleus, as illustrated here

neighboring positive ion cores and are therefore equally shared. However, when there is a covalent-type bond between two different atoms, the electrons are unequally shared because the two neighboring ion cores are different and hence have different electron-attracting abilities. The bond is no longer purely covalent but has some ionic character, because the shared electrons are more shifted toward one of the atoms. In this case a covalent bond has an ionic component and is generally called a polar bond. Many technologically important semiconductor materials, such as III–V compounds (e.g., GaAs, InSb, and so on), have polar covalent bonds. In GaAs, for example, the electrons in a covalent bond are closer to (i.e., more probably found near) the As ion core than the Ga ion core. This example is shown in Fig. 1.12.

In ceramic materials, the type of bonding may be covalent, ionic, or a mixture of the two. For example, silicon nitride (Si_3N_4), magnesia (MgO), and alumina (Al_2O_3) are all ceramics, but they have different types of bonding: Si_3N_4 has covalent, MgO has ionic, and Al_2O_3 has a mixture of ionic and covalent bondings. All three are brittle, have high melting temperatures, and are electrical insulators.

1.5.5 Metallic Bonds

Atoms in a metal have only a few valence electrons, which can be readily removed from their shells and become collectively shared by all the resultant ions. The valence electrons therefore become delocalized and form an electron gas, permeating the space between the ions, as depicted in Fig. 1.13. The attraction between the negative charge of this electron gas and the metal ions forms the bonding in a metal. However, the presence of this electron cloud also adds a repulsive force to the bonding. Nevertheless, overall, Fig. 1.7 is still valid except that the cohesive energy is now lower in absolute value compared to ionic and covalent bonds, i.e., it is easier

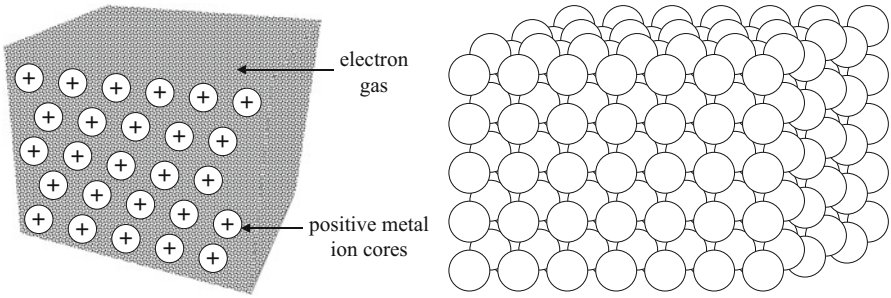


Fig. 1.13 Metallic bonding resulting from the attraction between the electron gas and the positive metal ions. The electrons are delocalized inside the volume between the atoms in the crystal

in many cases to "pull apart" metal regions, which explains why metals are usually malleable.

This metallic bond is nondirectional (isotropic). Consequently, metal ions try to get as close as possible, which leads to close-packed crystal structures with high coordination numbers, compared to covalently bonded solids. "Free" valence electrons in the electron gas can respond readily to an applied electric field and drift along the force of the field, which is the reason for the high electrical conductivity of metals. Furthermore, if there is a temperature gradient along a metal bar, the free electrons can also contribute to heat transfer from the hot to the cold regions. Metals therefore also have a good thermal conductivity.

1.5.6 Secondary Bonds

Since the atoms of inert elements (column VIII in the periodic table) have full shells and therefore cannot accept any extra electrons nor share any electrons, one might think that no bonding is possible between them. However, a solid form of argon does exist at temperatures below −189 °C, which means that there must be some type of bonding mechanism between the Ar atoms. However, the bond energy cannot be high since the melting temperature is so low.

A particular type of weak attraction that exists between neutral atoms and molecules involves the so-called dipolar and the van der Waals forces, which are the result of the electrostatic interaction between permanent or temporary electric dipoles in an atom or molecule. An electric dipole occurs whenever there is a separation between a negative and a positive charge of equal magnitude Q, as shown in Fig. 1.14a. A dipole moment is defined as a vector $\vec{p} = Q\,\vec{x}$, where $\vec{x}$ is a distance vector from the negative to the positive charge.

One might wonder how a neutral atom can have an electric dipole. We know that electrons are constantly moving in orbitals around the nucleus. As a result of this motion, the distribution of negative charges is never exactly centered on the nucleus, thus yielding a tiny, transient electric dipole. A dipole moment can also be a

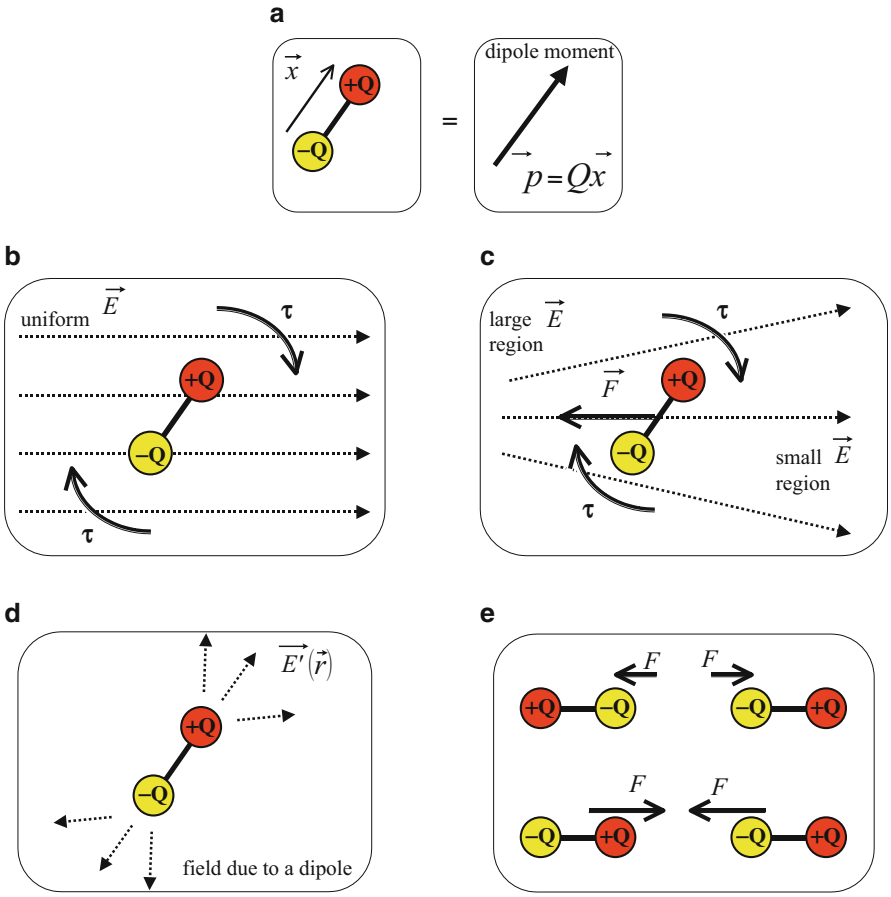


Fig. 1.14 Electric dipole moment and its properties. (**a**) A dipole is formed when two electrical charges with opposite signs and equal magnitude are separated by a distance. This creates a dipole moment. (**b**, **c**) A dipole can rotate and be translated in the presence of an electric field. (**d**) A dipole creates an electric field of its own, as a result of its two constituting electrical charges. (**e**) Dipoles can interact with each other because one will feel the electric field produced by the other

permanent feature of a molecular structure or induced by an external electric field. In the latter case, the atom or molecule in which a dipole moment appears is said to be polarized by the external electric field.

When an electric dipole is placed in an external electric field $\vec{E}$, it will experience both a torque $\boldsymbol{\tau}$ and a force $\vec{F}$ (unless the external electric field is uniform in space) as a result of the electrostatic forces exerted on each charge by the electric field, which is depicted in Fig. 1.14b, c. In a uniform field, the torque $\boldsymbol{\tau}$ will simply try to rotate the dipole to line up with the field, because the charges $+Q$ and $-Q$ experience similar magnitude forces in opposite directions. In a nonuniform field, the net force

F experienced by the dipole tries to move the dipole toward stronger field regions. This force will depend on both the orientation of the dipole and the gradient of the electric field.

Moreover, a dipole moment creates an electric field $\vec{E}'\left(\vec{r}\right)$ of its own around it as shown in Fig. 1.14d, just as a single charge does. Therefore, a dipole can interact with another dipole as shown in Fig. 1.14e. This interaction is also at the origin of the van der Waals force and the van der Waals bond. The van der Waals bond is the result of the attraction caused by the instantaneous dipole of one atom inducing a dipole in another atom. It occurs even when the atoms have no permanent (time averaged) dipole moment. This bond is very weak and its magnitude drops rapidly with distance R, namely, as $1/R^6$. Figure 1.7 is nevertheless still valid but with a much smaller cohesive energy. The bond energy of this type is at least an order of magnitude lower than that of a typical ionic, covalent, and metallic bonding. This is why inert elements such as Ne and Ar solidify at temperatures below 25 K (−248 °C) and 84 K (−189 °C), respectively.

In some solids, a van der Waals force may dominate in one direction, while an ionic and/or covalent bond dominates in another. Several solids may therefore have dominant cleavage planes perpendicular to the van der Waals force directions. Moreover, many solids that we say are mostly ionic or covalent may still have a very small percentage of van der Waals force present too. Graphite is a typical example. It is made up of stacks of sheets of carbon. In one sheet the carbon atoms are covalently bound. However, the sheets are held together only by van der Waals forces, and as a result the sheets slide easily over each other making graphite easily cleavable and very soft, properties put to good use in pencil lead.

There is a special class of bond called the hydrogen bond, in liquids and solids where the attraction between atoms or molecules appears through a shared proton. Figure 1.15 shows the hydrogen bond in the H_2O molecule. Such a molecule has a permanent dipole moment. Each proton in a molecule can form a bond with the oxygen in two other molecules. This dipole-dipole interaction keeps water molecules together in liquid water or solid ice.

The greater the energy of the bond is, the higher the melting temperature of the solid is. Similarly, stronger bonds lead to greater elastic moduli and smaller expansion coefficients.

1.6 Atomic Property Trends in the Periodic Table

1.6.1 The Periodic Table

As its name suggests, the periodic table of elements is organized based on the periodicity of the electronic structure in atoms. In the periodic table, all the elements in the same row make up a period (in this discussion "across a period" will mean from left to right), and all the elements in a column are a group. Elements in a group have the same valence shell configuration. The part of the

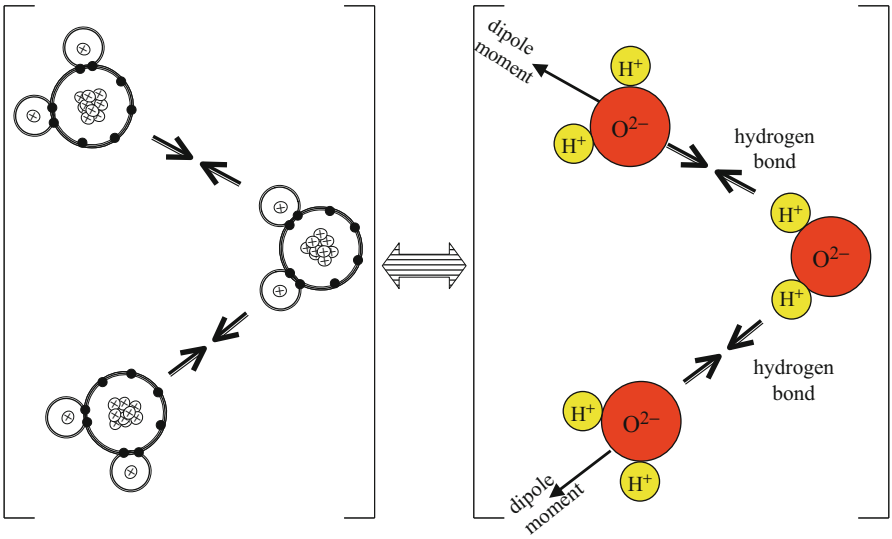


Fig. 1.15 The origin of hydrogen bonding between water molecules. A H_2O molecule has a net permanent dipole moment as a result of its lack of central symmetry. The H_2O molecules can therefore interact with one another. Attractions between the various dipole moments in water give rise to hydrogen bonding

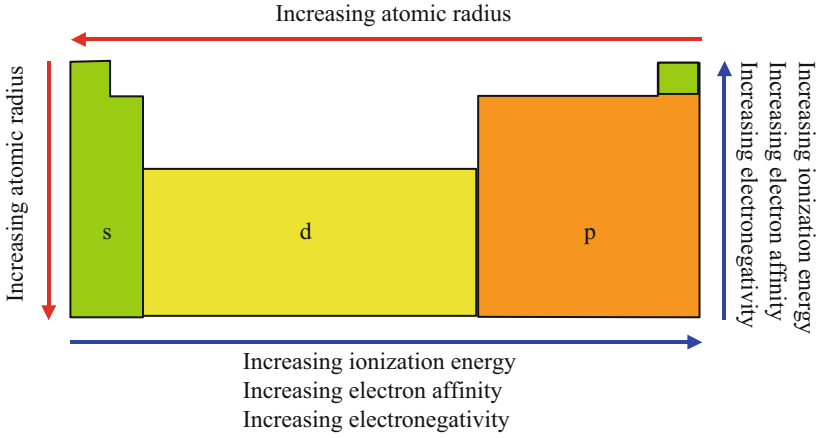


Fig. 1.16 Part of the periodic table with divisions indicating valence shells and a summary of atomic property trends

periodic table shown in Fig. 1.16 can be divided into three sections that indicate which orbitals (*s*, *p*, or *d*) are the valence shell. The *f*-orbital valence shell elements are omitted for simplicity.

The electron configuration of an atom (especially that of its valence shell) is a primary determinant of the atom's properties. As a result, the variation of atomic

properties across the table should reflect the "structure" of the periodic table. This can be seen in many of the basic atomic properties. The discussion here will focus on atomic and ionic radii, ionization energy, electron affinity, and electronegativity. The variation trends of these properties across a period (from left to right) and down a group are very good examples of the role of the interatomic electrical forces. The properties discussed here are determined by the interplay between nuclear attraction of electrons, electron-electron repulsions, and nuclear-charge screening.

1.6.2 Atomic and Ionic Radii

Since electrons in an atom are delocalized in the orbitals, not only does the orbital not have a well-defined boundary, but the whole atom also does not have a well-defined size. Typically, the atomic radius (a spherical shape is generally assumed) is instead defined by half the distance between the atoms in a chemical compound. This definition is oversimplified since different atoms form different types of bonds, but regardless, trends can still be observed.

The atomic radius decreases going across a "period" and increases going down a group. Going across a period, protons and electrons are being incrementally added. The dominant force originates from the increased nuclear charge attracting the electron clouds more strongly. Going down a group, the atomic radius increases because electrons are occupying larger orbitals corresponding to higher and higher principal quantum numbers.

Another important size is that of an element's ion compared to its neutral state. A positively ionized atom has lost an electron from the outermost (largest) shell, which reduces its size. Also, the loss of an electron reduces the electron-electron repulsions in the orbitals that would otherwise cause them to spread out over a larger space. A negative ion is larger than the neutral ion because the additional electron increases electron-electron repulsions. The change in size for ions can be very large. For example, the radius of Li changes from 1.52 Å to 0.76 Å when it loses an electron.

1.6.3 Ionization Energy

Ionization energy is defined as the energy required to remove an electron from an atom or ion, creating a more positive particle. In the ionization process, the highest energy, or outermost, electron is removed. The energy required to remove an electron from an atom in its ground state is called the first ionization energy. The energy required to remove a second electron is called the second ionization energy and so on. As the degree of ionization increases, so does the energy required. This is because it is increasingly more difficult to remove a negative charge from an increasingly positively charged ion. As the ion becomes more positive, it attracts any electrons around it more strongly because the effective nuclear charge they experience is larger. From the point of view of the orbital model, taking successive electrons from an atom requires reaching deeper into the atom to remove an electron

from the more tightly bound lower energy levels. The ionization energy always jumps by a large amount once all the valence electrons have been removed, and ionization from the full shell starts.

Going across a period, the first ionization energy increases due to increased nuclear attraction. This is like the trend for atomic radius. Going down a group, the first ionization energy decreases because the ionized electron is coming from orbitals with a higher principal quantum number. In these higher orbitals, the electron spends the majority of its time further from the nucleus and so the atom is easier to ionize.

1.6.4 Electron Affinity

The electron affinity is the potential energy change of the atom when an electron is added to a neutral, gaseous atom to form a negative ion. So the more negative the electron affinity, the more favorable the electron addition process is. Not all elements form stable negative ions, in which case the electron affinity is zero or even positive (energy is required to add an electron).

Of the properties discussed, electron affinity is the least well behaved because it has the most exceptions. It is also difficult to measure. There is a tendency toward increased electron affinity going left to right across a period. The overall trend across a period occurs because of increased nuclear attraction. The exceptions occur because, for certain electron configurations, the electron-electron repulsion force (not to be confused with screening) is stronger than the nuclear attraction. Exceptions also occur because those elements that have completely filled valence shells are particularly stable. Going down a group, the electron affinity should decrease since the electron is being added increasingly further away from the atom (i.e., less tightly bound and therefore closer in energy to a free electron). In reality, this trend is a very weak one as the affinities do not change significantly down most groups.

1.6.5 Electronegativity

Electronegativity is a measure of the ability of an atom in a molecule to attract shared bonding electrons. This property is different from the other ones presented here because it is not relevant for an isolated atom since it deals with shared electrons. A higher electronegativity means that the atom will attract bonded electrons to it more strongly. Electronegativity increases across a period and decreases down a group. The difference in electronegativity between bonding atoms determines whether the bond is covalent, ionic, or in between (polar covalent). For atoms with similar electronegativity, neither atom attracts the shared electron more strongly. This equal sharing is characteristic of a purely covalent bond. As the electronegativity difference increases, the shared electron will spend more time near the more electronegative atom. The unequal sharing results in a polar covalent bond, which in the extreme case of complete electron transfer becomes an ionic bond.

1.6.6 Summary of Trends

The different trends are summarized in Fig. 1.16. Appendix A.3 contains periodic tables that give the atomic radius, ionization energy, electron affinity, and electronegativity for all the elements. Understanding these trends allows one to understand properties not only of individual elements but also solid properties like lattice constants and semiconductor bandgaps. It is important to keep in mind that the trends discussed here are just generalizations, and exceptions do occur throughout the table. A more detailed discussion of these properties and the exceptions can be found in most general chemistry texts (see Further Reading section).

1.7 Introduction to Energy Bands

So far, we have considered the concepts associated with the formation of bonds between two atoms. Although these concepts are important issues in semiconductor materials, they cannot explain a number of semiconductor properties. It is necessary to have more detailed information on the energies and the motion of electrons in a crystal, as well as understand the electron collision events against imperfections of different kinds. To do so, we must first introduce the concept of energy bands. The formation of energy bands will be discussed in more detail in Chap. 5 using a quantum mechanical formalism. However, for the moment, energy bands can be conceptually understood by considering a simple example.

The electronic configuration in an isolated Si atom is such that 10 of its 14 electrons are tightly bound to the nucleus and play no significant role in the interaction of the Si atom with its environment, under all familiar solid-state device conditions. By contrast, the remaining four valence electrons are rather weakly bound and occupy four of the eight allowed energy states immediately above the last core level. For a group of N isolated Si atoms, i.e., far enough apart so that they are not interacting with one another, the electronic energy states of their valence electrons are all identical.

When these N atoms are brought into close proximity, to form crystalline Si, for example, the energy levels for the outer electrons are modified as shown in Fig. 1.17b. Exactly half of the allowed states become depressed in energy (bonding states) and half increase in energy (antibonding states). Moreover, this perturbation does not leave the energy levels sharply defined but spread them into bands instead. Two bands of allowed electronic energy states are thus formed, as shown in Fig. 1.17b, which are separated by an energy gap, i.e., an energy region forbidden for electrons where there is no allowed electronic energy state.

At very low temperatures, the electrons fill the low-energy band first. The band below the bandgap in energy is called the valence band. The band above the bandgap, which is not completely filled and in most cases completely empty, is called the conduction band. The energy gap between the highest energy level in the valence band and the lowest energy level in the conduction band is called the bandgap.

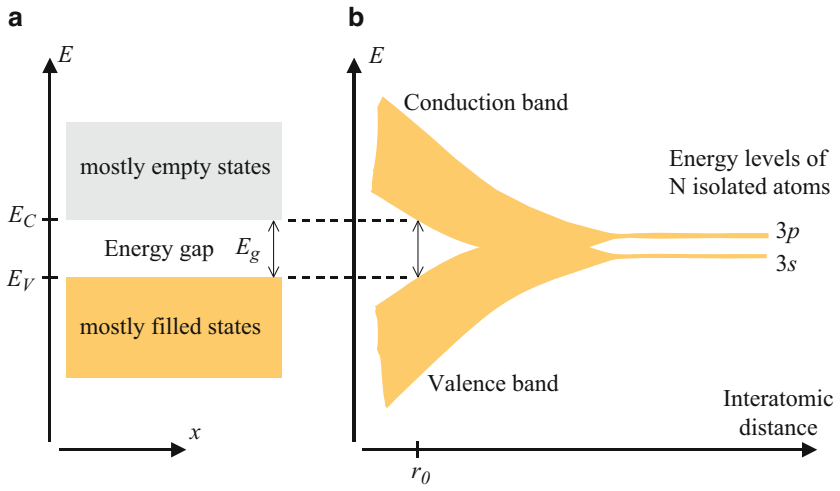


Fig. 1.17 Illustration of the formation of energy bands in a Si crystal. A system of N isolated Si atoms has discrete allowed energy levels, all located at the energies of the 3*s* and 3*p* orbitals of an isolated Si atom. When the atoms come into close proximity, the energy levels are modified as shown in the figure, as a result of the interaction between the atoms. The allowed energy levels start to form energy bands

It should be noted that the band electrons in crystalline silicon are not tied to or associated with any one particular atom. On average, one will typically find four valence electrons being shared between any given Si atom and its four nearest neighbors (as in the bonding model). However, the identity of the shared electrons changes as a function of time, with the electrons moving around from point to point in the crystal. In other words, the allowed electronic states or bands are no longer atomic states but are associated with the crystal as a whole, independent of the point examined in a perfect crystal. An electron sees the same energy states wherever it is in the crystal.

We can therefore say that, for a perfect crystal under equilibrium conditions, a plot of the allowed electron energies versus distance along any preselected crystalline direction (x) is as shown in Fig. 1.17a. This plot is the basic energy band model. E_C introduced in Fig. 1.17a is the lowest possible conduction band energy, E_V is the highest possible valence band energy, and $E_g = E_C – E_V$ is the bandgap. A more detailed consideration of the bands and electron states will be given in Chap. 5.

The energy band and the bandgap concepts are at the heart of semiconductor physics. As the name implies, a semiconductor has an electrical conductivity in between that of a metal and an insulator. Also, in a semiconductor the electrical conductivity can be varied by changing the structural properties of the semiconductor, changing the temperature, or applying external fields. These properties are a direct consequence of the energy band structure. Understanding and utilizing these properties of semiconductors is the goal of this book.

1.8 Summary

In this chapter, the electronic structure of atoms and its implications on the bonding and the formation of energy bands in solids have been presented. Early experiments conducted on even the simplest atom that of hydrogen showed that classical mechanics was insufficient and that a new theory, called wave or quantum mechanics, was necessary in order to understand the observed physical phenomena.

The notion of electron density function and the Bohr radius have been introduced. The concepts of atomic orbitals and quantum numbers to identify the allowed discrete energy levels for electrons in an atom have been discussed. The nature of the bonding between atoms in a solid, be it ionic, covalent, mixed, metallic, or secondary, has been described by taking into account the interaction of electrons in the higher energy levels in the atoms in presence. Finally, the formation of energy bands and the concept of conduction and valence bands have been introduced through the interaction of multiple atoms.

Problems

1. The size of an atom is approximately 10^{-8} cm. To locate an electron within the atom, one should use electromagnetic radiation of wavelength not longer than 10^{-9} cm. What is the energy of the photon with such a wavelength (in eV)?
2. Using the Rydberg formula, calculate the wavelength and energy of the photons emitted in the Lyman series for electrons originally in the orbits $n = 2$, 3, and 4. Express your results in cm, eV, and J. In which region of the electromagnetic spectrum are these emissions?
3. What are the radii of the orbits and the linear velocities of the electrons when they are in the $n = 1$ and $n = 2$ orbits of the hydrogen atom?
4. Using Bohr's model, deduce an analytical expression for the Rydberg constant as a function of universal constants.
5. The He^+ ion is a one-electron system similar to hydrogen, except that it has two protons. Calculate the wavelength of the longest wavelength line in each of the first three spectroscopic series ($n = 1$, 2, 3).
6. The human eye is more sensitive to the yellow-green part of the visible spectrum because this is where the irradiance of the sun is maximum. Since the sun can be considered as a blackbody with a temperature of approximately 5800 K, use Planck's relation for the irradiance of a blackbody $I(\lambda) = \frac{2\pi hc^2}{\lambda^5}\left\{\frac{1}{e^{\frac{hc}{\lambda k_b T}}-1}\right\}$ to find the wavelength of the maximum of the sun irradiance. You will come out with a very simple relation between the peak of the irradiance (λ_{peak}) and T, which is called Wien's relation. In Planck's relation above, h, c, λ, k_b, and T are, respectively, Planck's constant, the velocity of light in vacuum, the wavelength, Boltzmann's constant, and the absolute temperature. You will need the

following solution for the equation $x = 5(1-e^{-x})$, $x = 4.965$. Then use Wien's relation to estimate λ_{peak} for a human body.

7. Since an electron on a circular orbit around a proton has a centripetal acceleration, it should radiate energy according to the Larmor relation $dE/dt = -2/3\ (q^2/4\pi\varepsilon_0)\ (a^2/c^3)$ where q, a, ε_0, and c are, respectively, the electron charge, its acceleration, the vacuum permittivity, and the velocity of light in vacuum. Therefore, in classical mechanics, it should spiral and crash on the nucleus. How long would this decay take, supposing that the size of the initial orbit is 10^{-10} m and the nucleus is a point charge (i.e., radius = 0)?
8. What is Hund's rule? Show how it is used to specify in detail the electron configurations of the elements from Li to Ne.
9. What is the full electronic configuration of Li? Since the ionization energy of Li is 5.39 eV, how much is the effective nuclear charge? What can you say about the screening of the other electrons?
10. Calculate the total coulombic potential energy of a Na^+ in a NaCl crystal by considering only up to the fourth nearest neighbors of Na^+. The coulombic potential energy for two ions of opposite charges separated by a distance r is given by:

$$E(r) = -\frac{q^2}{4\pi\varepsilon_0 r} \quad (q > 0).$$

11. The interaction energy between Na^+ and Cl^- ions in the NaCl crystal can be written as

$$E(r) = -\frac{4.03 \times 10^{-28}}{r} + \frac{6.97 \times 10^{-96}}{r^8}$$

where the energy is given in joules per ion pair and the interionic separation r is in meters. The numerator unit of the first term is J·m and the second term is J·m^8. Calculate the binding energy and the equilibrium separation between the Na^+ and Cl^- ions.

12. Consider the van der Waals bonding in solid argon. The potential energy as a function of interatomic separation can generally be modeled by the Lennard-Jones 6–12 potential energy curve, that is, $E(r) = -Ar^{-6} + Br^{-12}$ where A and B are constants. Given that $A = 1.037 \times 10^{-77}$ J·m^6 and $B = 1.616 \times 10^{-134}$ J·m^{12}, calculate the bond length and bond energy (in eV) for solid argon.
13. Which group of the periodic table would you expect to have the largest electron affinities?
14. Which atom has the higher ionization energy, zinc or gallium? Explain.
15. Arrange the following groups of atoms in order of increasing size (without resorting to the tables in the appendices).
    a. Li, Na, K
    b. P, S, Cl
    c. In, Sn, Tl
    d. Sb, S, Cl, F

16. Based on the electronegativities given in Fig. A.1 in Appendix A.3, what groups of elements would you expect to form ionic compounds? Is this consistent with reality?
17. Why do none of the noble or inert gases (elements in the rightmost group) have electron affinity values listed in Appendix A.3 Fig. A.?

Further Reading

Atkins PW (1983) Molecular quantum mechanics. Oxford University Press, New York
Cohen MM (1972) Introduction to the quantum theory of semiconductors. Gordon and Breach, New York
Ferry DK (1991) Semiconductors. Macmillan, New York
Kasap SO (1997) Principles of engineering materials and devices. McGraw-Hill, New York
Kittel C (1976) Introduction to solid state physics. Wiley, New York
Pierret RF (1989) Semiconductor fundamentals. Addison-Wesley, Reading, MA
Pierret RF (1989) Advanced semiconductor fundamentals. Addison-Wesley, Reading, MA
Yu PY, Cardona M (1999) Fundamentals of semiconductors: physics and materials properties. Springer, New York
Ziman JM (1998) Principles of the theory of solids. Cambridge University Press, Cambridge
Zumdahl SS, Zumdahl SA (2003) Chemistry. Houghton Mifflin Company, Boston

The Carbon Atom

2

2.1 Introduction: The Carbon Atom

Carbon is the 15th most abundant element in the Earth's crust and the fourth most abundant element in the universe by mass after hydrogen, helium, and oxygen. It is present in all known life forms. In the human body, carbon is the second most abundant element by mass (about 18.5%) after oxygen. This abundance, together with the unique diversity of organic compounds and their unusual polymer-forming ability at the temperatures commonly encountered on Earth, makes this element the chemical basis of all known life [Demarchi, Falkowski, Gruber]. More precisely, the carbon atom forms a number of components comparable with the total addition of all the other elements of the periodic table in combination with each other. In particular, we know more than 1 million organic components formed with only carbon and hydrogen.

As a member of group 14 in the periodic table (see Fig. 2.1), carbon is nonmetallic and tetravalent – making four electrons available to form covalent chemical bonds. This property is primordial to describe the resulting characteristics of carbon components and explain why it is essential to life. Indeed, carbon forms strong single bonds to itself that are strong enough to resist most of reactions at ambient conditions giving the carbon the possibility to form long chains of atoms, which are essential for many compounds in the living cell such as DNA.

2.1.1 Isotopes of Carbon Atom (Fig. 2.2)

There are in total 15 known isotopes of carbon, with varying atomic mass varying from 8 to 22 (^{8}C to ^{22}C), which differ only in their number of neutrons. ^{12}C and ^{13}C are the only stable isotopes, while the others are radioactive. The most stable radioisotope is ^{14}C decaying with a half-life of about 5730 years while all other isotopes of carbon have half-lives less than 20 s. In this textbook, and unless

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_2

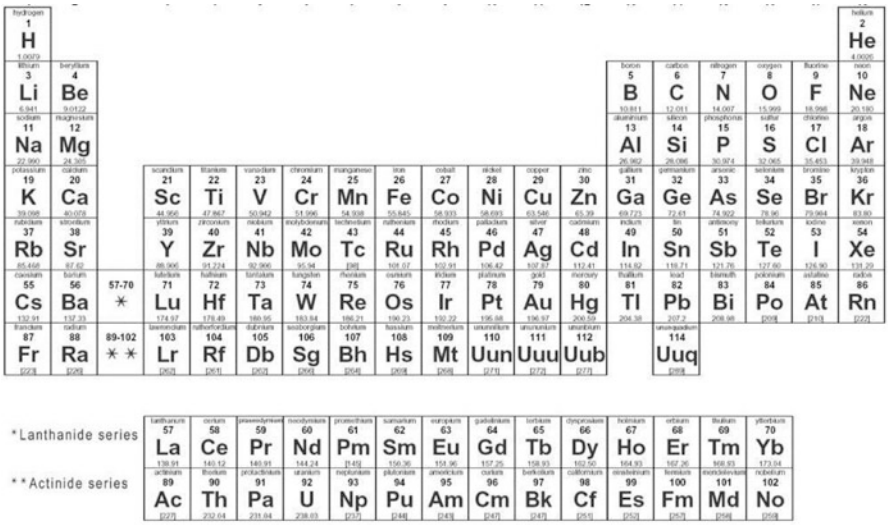


Fig. 2.1 The periodic table, carbon is in group 14

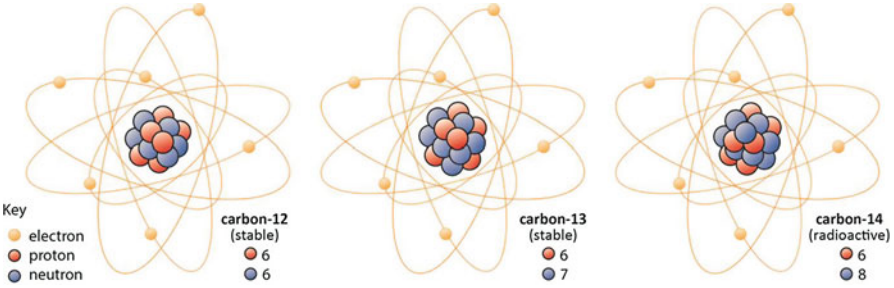


Fig. 2.2 The three natural isotopes of carbon. The two left sketches are ^{12}C and ^{13}C, the only two stable isotopes. ^{14}C is the most stable radioisotope

mentioned, the properties given and calculations are assumed to be done for ^{12}C isotope, which is the most common isotope that can be found in the nature, since measurements by mass spectroscopy show that about 99% of all carbon atoms are in ^{12}C isotope. Note that ^{12}C single atom's weight is used in order to define the unified atomic mass unit or Dalton. More precisely, this unit is defined as one twelfth of the mass of an unbound atom of ^{12}C and has a value of $1.660538921 \times 10^{-27}$ kg.

Comparison between isotopes of carbon is given in Table 2.1.

2.1.2 Electronic Configuration

The electronic configuration of carbon is $[1s^2]\ 2s^2\ 2p^2$. This means the $n = 1$ shell is full with two electrons and the $n = 2$ shell has two electrons in the s-state (full) and

Table 2.1 Some fundamental properties of Carbon

	Z	N	Mass (u)	Half-life (y)	Decay mode	Nuclear spin
^{12}C	6	6	12	Stable		0+
^{13}C	6	7	13.003355	Stable		½–
^{14}C	6	8	14.003242	5730	β–	0+

[Audi, De Laeter, Wieser]
Atomic mass: 6
Atomic weight: ^{12}C ^{is}: 12.0107 ± 0.0008 u ($1u = 1.66053892 \times 10^{-27}$ kilograms)
Van der Walls radius: 170 pm

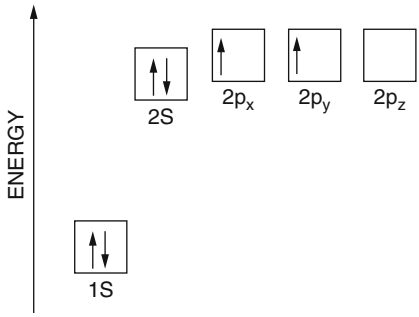


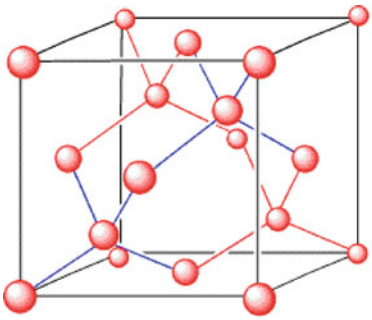
Fig. 2.3 The energy level structure of the carbon atom. Hund's rule (see [Razeghi] page 55) gives a triplet ground state: when there is space in the shell, electron/electron repulsion favors the triplet state

two electrons in the p-state which can accommodate up to six electrons. A reminder of how electronic configuration is obtained can be found in (Appendix 1 – Atomic Orbital).

The surface shell or n shell of an atom determines its reactivity. For carbon atom, this surface shell is $n = 2$ and its study explains the properties of bindings of the atom. In order to optimize bonding with other carbon atoms and chemicals, one electron from the $2s^2$ orbital in the surface $n = 2$ shell can be promoted to become a p orbital electron, so that now we have $2s2p^3$ instead of $2s^2\ 2p^2$. As shown in Fig. 2.3, the 2p orbital energies are higher than the energy of 2s orbital; thus this promotion costs energy, which has to be recouped by the bond it forms.

In a three-dimensional bond like in diamond (see Fig. 2.4), the 2s electron combines with the three 2p orbitals (Appendix 1 and Fig. 2.4) to form four directed bonds in a tetrahedral arrangement which can now optimally overlap with the similarly formed wavefunction of neighboring carbons a shown in Fig. 2.4. This bonding optimizes the overlap of negative with positive charge and is called sp^3. It lowers the energy much more than the price paid to lift the s to p from $2s^2\ 2p^2$ to $2s2p^3$.

But this is not all; carbon can also link in a planar arrangement forming only three sp^3 bonds in a plane as in graphene and leaving on p orbital standing. Finally, the sp^3 can also link in a linear configuration forming only two linear bonds with two neighboring atoms, thus leaving 2p orbital standing, and this will be shown later.



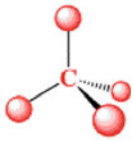


Fig. 2.4 The carbon tetrahedral sp^3 bonds and the formation of the covalent three-dimensional sp^3 covalent diamond lattice; see FSSE [Razeghi] Chap. 1

Table 2.2 Binding energy and other properties of bonds involving Carbon

Bond	C–C	C=C	C≡C	C–Si	C–Ge	C–Sn	C–Pb	C–N	C=N	C≡N
D(kJ/mol)	346	602	835	318	238	192	130	305	615	887
R(pm)	154	134	120	185	195	206	230	147	129	116
C–P	C–O	C=O	C ≡ O	C–B	C–S	C=S	C–F	C–Cl	C–Br	C–I
264	358	799	1072	356	272	573	485	327	285	213
184	143	120	113		182	160	135	177	194	214

[Reference Cottrell, Darwent]

In the sp^3 tetrahedral bond as in diamond Fig. 2.4, the bond length called the covalent radius is only 77 pm. It is 73 pm in graphene and 69 pm in a linear polymer. The first ionization energy of carbon, i.e., the energy required to remove an electron from carbon atom in the gas phase, is first 1086.5 kJ mol^{-1}.

2.1.3 Binding Energies

This table provides useful constants concerning covalent bindings involving the carbon atoms: the bonding energy D, which can be interpreted as the energy necessary to break the bond, and the bonding distance between the two atoms, R (Table 2.2).

2.2 Covalent Bonding Between Carbon Atoms

In the carbon ground state, the *p-shell is doubly occupied* $2s^2 2p^2$ *and can acquire an* electron from the filled s-shell below to form covalent bonds which lowers the energy sufficiently to pay the price for the initial uplift "s to p." This is similar to what happens in other materials such as Si, $3s^2 3p^2$, and Ge, $4s^2 4p^2$, except that with

carbon, the bonding need not be tetrahedral, i.e., to four neighbors involving four sp^3 directional covalent bonds as in Si, for example, or diamond. The carbon bonding can involve three covalent bonds sp^2 planar with one extra p orbital perpendicular, or indeed it can be an s-p bond and linear. We will investigate the consequences of the different bonding arrangements as we develop the chapter.

A *carbon-carbon bond* is a covalent bond between two carbon atoms. The most common form is the single bond: a bond composed of two electrons, one from each of the two atoms. The carbon-carbon single bond is a sigma bond and is said to be formed between one hybridized orbital from each of the carbon atoms. In ethane C_2H_6, the orbitals are sp^3 hybridized orbitals, but single bonds formed between carbon atoms with other hybridizations do occur (e.g., sp^2 to sp^2). In fact, the carbon atoms in the single bond need not be of the same hybridization. Carbon atoms can also form double bonds in compounds called *alkenes* (see Fig. 2.5). *Alkenes* are a class of hydrocarbons that contain only carbon and hydrogens. They are unsaturated compounds that contain at least one carbon-to-carbon double bond. Another term that is often used to describe *alkenes* is olefins or triple bonds in compounds called alkynes.

A double bond is formed with a sp^2 hybridized orbital and a p orbital that isn't involved in the hybridization. In *alkynes* a triple bond is formed with a sp hybridized orbital and two p orbitals from each atom (see Fig. 2.6). The use of the p orbitals forms a pi bond.

Carbon is one of the few elements that can form long chains of its own atoms, a property called catenation. This coupled with the strength of the carbon-carbon bond gives rise to an enormous number of molecular forms, many of which are important structural elements of life, so carbon compounds have their own field of study: organic chemistry.

Branching is also common in C-C skeletons. Different carbon atoms can be identified with respect to the number of carbon neighbors.

- *Primary carbon atom*: one carbon neighbor
- *Secondary carbon atom*: two carbon neighbors
- *Tertiary carbon atom*: three carbon neighbors
- *Quaternary carbon atom*: four carbon neighbors

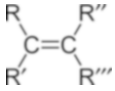


Fig. 2.5 Alkene flat, the R R″ are any side groups that attach to C with a single bond and fit into space

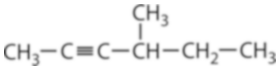


Fig. 2.6 4-methyl-2-hexyne molecule, example of a triple carbon bond

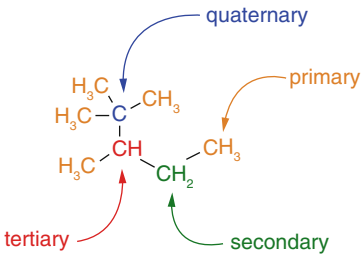


Fig. 2.7 Trimethylpentane example of the versatile way carbon can bond to itself and other compounds

It is this versatility and flexibility shown in Fig. 2.7, its ability to bond in many different configurations, which makes carbon such a special and unique material and essential component to life.

In the next section, we will be looking at the structure of the most important currently known carbon allotropes. This is also the main focus of this review. The vast field of carbon chemistry and physics in conjunction with carbon compounds cannot be covered by this review; however when a particular point of scientific interest arises, it will be mentioned and the reference provided. The emphasis of later chapters as we proceed is on pure carbon allotropes. We will discover their electronic and optical properties and investigate what makes them special.

Already now if we ask the question what is the mystery of carbon? Maybe the answer is it is the mystery of the carbon-carbon bond, sp, and sp^2 and sp^3 bond.

2.3 Carbon Allotropes

When an element of the periodic table exists in more than one crystalline forms, those forms are called allotropes.

There are several allotropes of carbon of which the best known are the three crystalline structures *graphite, diamond*, and *lonsdaleite*. The physical properties of carbon vary dramatically with the allotropic form, which is why recent researches in materials are mainly focusing in the study of these forms. For example, diamond is highly transparent, while graphite is opaque and black. Diamond is the hardest naturally occurring material known, while graphite is soft enough to form a streak on paper. Diamond has a very low electrical conductivity, while graphite is a very good conductor. Under normal conditions, diamond, carbon nanotubes, and graphene have the highest thermal conductivities of all known materials. High thermal conductivity is crucial in the field of power electronics and also in the area of big machines for computational science and the trillion dollar personal computer and laptop industries (Fig. 2.8).

Some Allotropes of Carbon

(a) Diamond, tetrahedral bonding sp^3 (discussed in Chap. 3).
(b) Graphite, two-dimensional sp^2 bonding, and van der Waals bonded layered structure (van der Waals bonding is a weak bonding via mutually induced

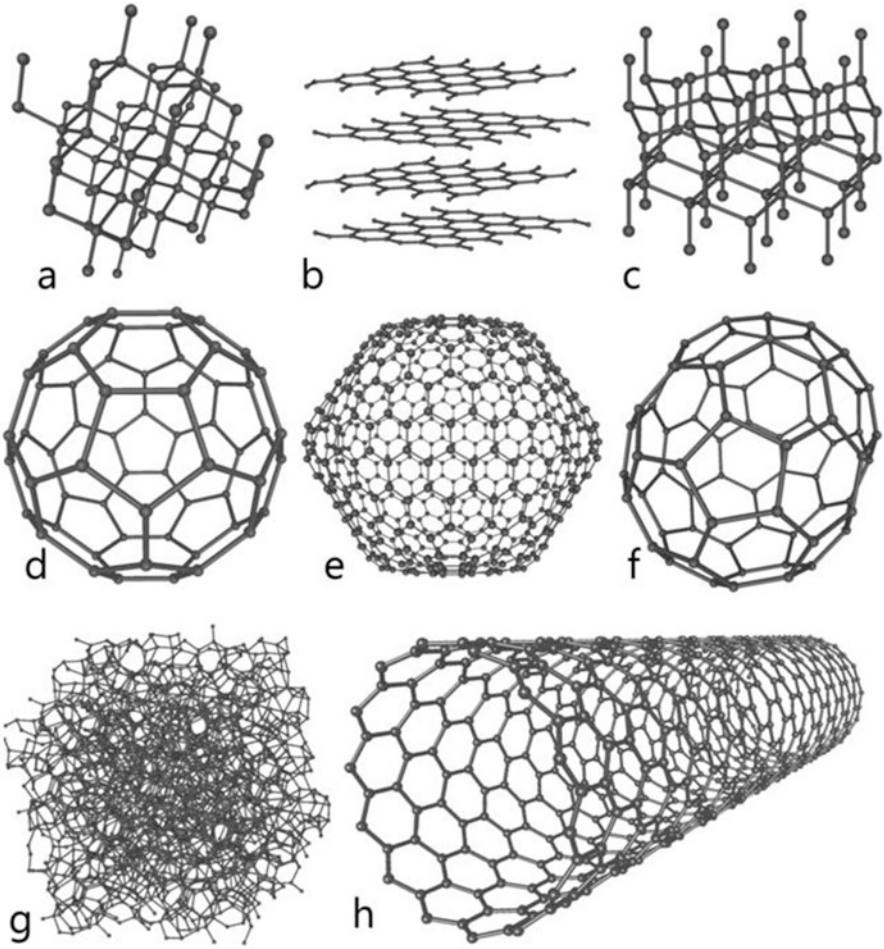


Fig. 2.8 Eight of the allotropes (different molecular configurations) that pure carbon can take: (**a**) diamond, (**b**) graphite, (**c**) lonsdaleite, (**d**) C60 (buckminsterfullerene), (**e**) C540 (see fullerene), (**f**) C70 (see fullerene), (**g**) amorphous carbon, (**h**) single-walled *carbon nanotube*

dipole-dipole coupling). The two-dimensional layers can be peeled off to give grapheme. Full discussion is in Chap. 4.

(c) *Lonsdaleite* (named in honor of Kathleen Lonsdale), also called *hexagonal diamond* in reference to the crystal structure, is an allotrope of carbon with a hexagonal lattice. In nature, it forms when meteorites containing graphite strike the Earth. It is translucent, brownish-yellow, and has an index of refraction of 2.40–2.41 and a specific gravity of 3.2–3.3.

(d–f) Fullerenes, C60 ball formed by bonding 60 atoms in a sphere C540, C70 (see Chap. 4).

Table 2.3 A comparison of graphite, diamond, and other forms of carbon

Allotrope	Hybridization	Structure	Existence
Graphite	sp^2	Crystal, 2D hexagonal stacked	Natural
Graphene	sp^2	Crystal, hexagonal stacked	Natural or synthetic (monolayer)
Diamond	sp^3	Crystal, cubic	Natural
Lonsdaleite	sp^3	Crystal, 3D hexagonal	Natural
Fullerene	sp^2	Cluster	Synthetic
Nanotube	sp^2	With single, double, or multiple walls	Synthetic
Amorphous carbon	sp^2–sp^3	No crystalline structure or aggregate of crystals	Natural

Table 2.4 Comparison of graphite and diamond

	Graphite	Diamond
Mechanical hardness	Graphite is one of the softest materials known	Synthetic nanocrystalline diamond is the hardest material known
Lubricant properties	Graphite is very good lubricant displaying super lubricity	Diamond is the ultimate abrasive
Electrical conduction	Graphite is a conductor of electricity	Diamond is an excellent electrical insulator and has the highest breakdown electric field of any known material
Thermal conduction	Some forms of graphite are used for thermal insulation but some other forms are good thermal conductors	Diamond is the best known naturally occurring thermal conductor
Optical transparency	Graphite is opaque	Diamond is highly transparent
Lattice structure	Graphite crystallizes in the hexagonal system	Diamond crystallizes in the cubic system

(g) Amorphous carbon, dangling bonds can be saturated by hydrogen, for example.
(h) Carbon nanotube, tubelike bonding of pure carbon discussed in detail in Chap. 5.

This table compares some carbon allotropes and their electronic configuration and structural properties (Tables 2.3 and 2.4):

This table which only compares some basic properties of graphite and diamond gives us an idea of how the structural differences of a crystal composed of the same atom can give rise to dramatic consequences of the crystal's properties. The exact reasons of some of these properties remain today unknown, and one may imagine that newer crystalline structures and those yet to be discovered can give even more astonishing results, example being graphene, which is today a subject of intensive research. The following chapters of this textbook will try to analyze the difference of these structures and properties.

Table 2.5 Numerical values of physical properties in different allotropes of carbon

	Density (g.cm-3)	Molar heat capacity (J.mol-1.K-1)	Thermal conductivity (W.m-1.K-1)	Mohs hardness	Electrical resistivity (Ω.m)
Graphite	2.267	8.517	119–165	1–2	10^{-4}
Diamond	3.515	6.155	900–2300	10	$10^{11-}10^{18}$

The following Table 2.5 gives some numerical values of these properties, showing the large range of values that can be expected from allotropes of carbon.

2.4 Carbon Fullerenes

2.4.1 Buckyballs

While the possibility of a stable closed-cage molecular structure for carbon was first suggested in 1970, the existence of fullerenes was not verified experimentally until 15 years later [Kroto et al.]. Laser ablation of a graphite target was used to create carbon clusters. Mass spectra of the resultant vapor revealed the synthesis of molecules in two main groups – rings consisting of 10–30 atoms and larger molecules with predominantly 60 and 70 member atoms. Researchers soon predicted that these high-mass molecules possessed a closed-cage configuration. The name "fullerene" was coined after R. Buckminster Fuller, an architect renowned for his construction of geodesic domes resembling the structure of these molecules (Fig. 2.10). The fullerene lattice is similar to the hexagonal graphite lattice in that it consists of a two-dimensional surface. To create large curvature in a graphene sheet, the substitution of pentagons for hexagons is required. Geometrically, there are multiple arrangements that form a closed structure, but Euler's theory for polyhedra dictates that exactly 12 faces of the cage must be pentagonal, with any additional number of hexagonal faces. Thus, the smallest possible fullerene (C20) is composed solely of 12 pentagons. But the curvature induced by the pentagons comes with a price in the form of strain energy (the *sp*2 bonds are bent out-of-plane, resulting in significant *sp*3 character). This penalty is minimized by separating the pentagons in the lattice as much as possible. The smallest fullerene in which no two pentagons are adjacent is C60. The structural stability of C60 makes it the most abundant product of any fullerene growth process, typically 3–6 times more likely than the next most abundant product, C70.

The C60 molecules (often referred to as "buckyballs") are composed of 12 pentagonal and 20 hexagonal faces in a soccer-ball arrangement. The carbon bonds come in two varieties: single bonds along the 60 pentagonal edges, which measure 1.46 Å in length, and 30 electron-rich double bonds between adjacent hexagons, which are 1.40 Å in length. The mean molecular diameter as measured with NMR is 7.10 Å, consistent with the expected geometrical diameter of 7.09 Å when considering the atoms as points.

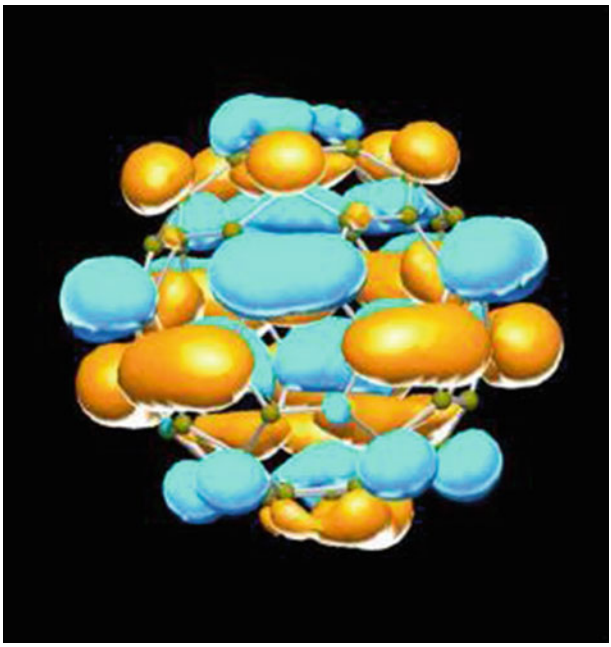


Fig. 2.9 Diagram of the wavefunction for the t_{1u} ($l = 5$) molecular orbital calculated using first-principle molecular dynamics method. Colors denote the sign of the wavefunction from Hornbaker, Thesis Univ. of Illinois. This is the lowest unoccupied molecular orbital in C_{60}. Its odd parity and threefold degeneracy make it very similar in character to atomic p orbitals

The electronic structure of C60 can be accurately approximated by considering its icosahedral symmetry. Each carbon atom contributes four valence electrons to the molecular structure. The σ-bonded *sp*2 electrons can be safely neglected as core-level molecular states, leaving 60 radially oriented *pz* orbital electrons to form the valence states. The irreducible representations of the icosahedral point group are used to determine the appropriate molecular orbital eigenfunctions. The spherical shape of C60 suggests an approximation of these molecular orbitals based on spherical harmonics *Y_l,m*. Since each angular momentum state *l* can accommodate $2(2l + 1)$ electrons, the first 50 electrons completely fill all states up to $l = 4$, leaving 10 electrons for the 22, $l = 5$ states.

Relative energy splittings within each level can be determined using a variety of computational methods. The lowest energy $l = 5$ states belong to the fivefold degenerate *hu* representation. These states account for the remaining ten electrons and are the highest occupied molecular orbitals (HOMOs) in the ground state. The lowest unoccupied molecular orbitals 5 (Fig. 2.9).

(LUMOs) are *t*1*u* states (Fig. 2.10), which are experimentally observed to reside ~1.9 eV above the *hu* levels in energy [Gunnarson]. The *t1u* molecular orbitals have the character of atomic *p* orbitals in that they are threefold degenerate and transform into one another under rotations about the [111] axis. In contrast, the *hu* orbitals have transformation properties resembling atomic *d* orbitals.

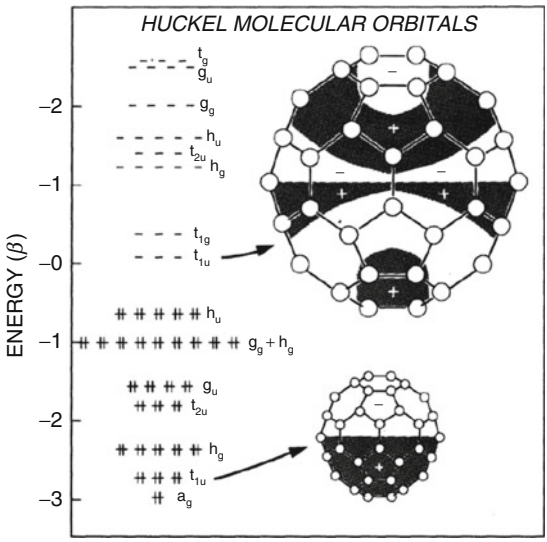


Fig. 2.10 Energy-level diagram of Hueckel π-molecular orbitals for C60, Ground state. Hueckel molecular orbitals and schematic illustration of some of the electronic wavefunction from Forro and Milhaly (2001) and ref. therein Nr 55, Haddon

In bulk, C60 forms a molecular solid with a face-centered cubic crystal structure at room temperature held together by van der Waals attraction. The nearest-neighbor distance is 10.02 Å, with an intermolecular separation (2.92 Å) similar to the spacing between layers in graphite (3.35 Å). Due to the relatively weak nature of van der Waals interactions, the constituent molecules rotate freely at room temperature. As the temperature is lowered below 260 K, rotations begin to freeze out and the buckyballs orient themselves relative to one another, leading to a lowering of the crystal symmetry to that of a simple cubic structure. The electronic structure of the solid is composed of bands derived from the molecular orbitals of the individual buckyballs. The undoped solid is a semiconductor with a 1.5 eV bandgap between the *hu*-derived valence band and the *t*1*u*-derived conduction band, which possess fairly narrow bandwidths of only ~0.4 V. Doping of C60 solids with alkali metal, alkali earth, or other elements can significantly change the conduction properties and in some cases even result in the onset of superconductivity.

The "spherical" symmetry leads to a high degree of level degeneracy which can be seen in Fig. 2.10 with a bandgap of about 1.8 eV.

2.5 Graphene and Nanotubes

Going back to Fig. 2.8 in the previous section, the next task is to obtain the one-dimensional band structure of an armchair nanotube (NT) (see Fig. 2.11) using the two-dimensional band structure of graphene. The latter has to be supplemented with periodic boundary conditions. The armchair NT is obtained by cutting out a slice from the graphene sheet parallel to the x axis. The slice has a width

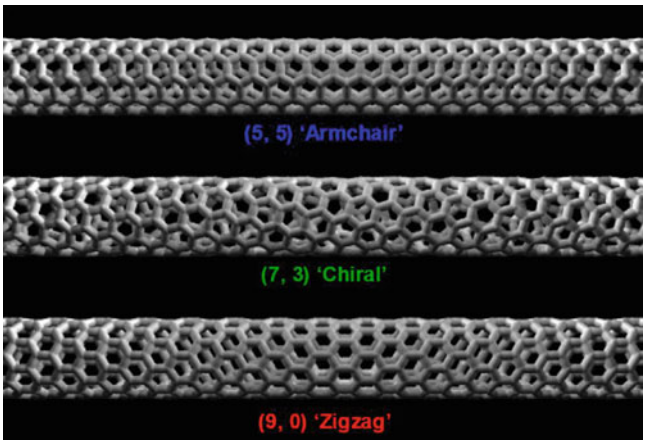


Fig. 2.11 The three structures of single-walled NT (n_1, n_2) and the wrapping vectors

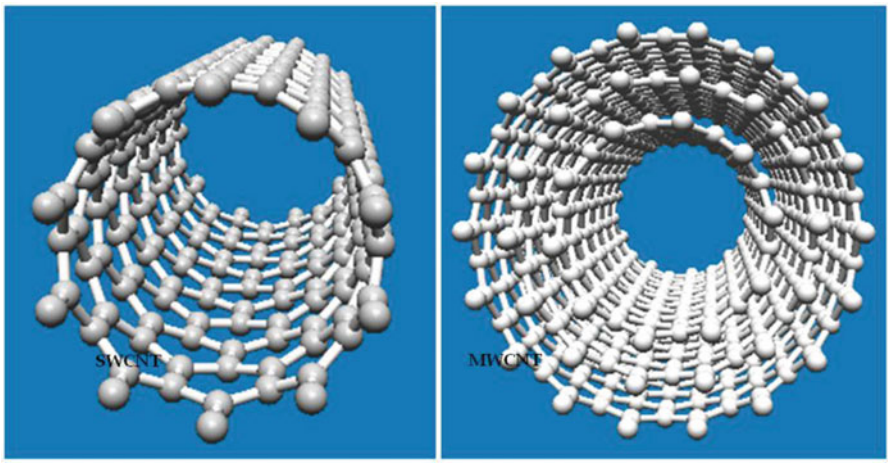


Fig. 2.12 SWNT left and MWNT (multiwalled) right; see also multiwalled carbon nanotubes (MWCNT): production, analysis, and application by AZoNano SouthWest NanoTechnologies (SWeNT)

w which can be expressed as the length of the so-called wrapping vector w orientated perpendicular to the tube axis. For an armchair NT, the wrapping vector is of the form $\vec{w} = N\left(\vec{a}_1 + \vec{a}_2\right)$ in general $n_1a_1 + n_2a_2$ where N is an integer. Usually this is also denoted as an (N, N) tube because the wrapping vector is equal N times a_1 plus N Times a_2. Due to the periodic boundary condition along the y-direction, the wavevector component k_y is quantized (Fig. 2.12).

The band structure of CNT is shown in Fig. 2.16 in the next section. Graphene is treated in more detail in Chap. 5.

2.6 Definition of Bonding Energy and Energy Bands

When two atoms with single orbitals such as two hydrogen atoms are brought close together, what happens? We have already touched on this question in the last chapter. Let us consider it again here and then develop it in a more rigorous platform in Chap. 5.

For each of the two hydrogen atoms, the lowest energy orbital is the 1s orbital and the potential energy corresponding decays exponentially in space so that at large separation almost no interaction between orbital takes place. Thus, the two atoms are symmetric and the energy level of the bonded electron is the same. As the second atom approaches to within the decay radius (1A°), the two orbitals start overlapping and the electrons can hop on to the other atom and can return. This redistributes the charge and gives rise to new energy levels. If the strength of this coupling energy is denoted by t_{12}, a simple calculation gives us the new levels as the bonding states with energy $2E_s$-2 t_{12} where E_s is the energy of the s-orbital and antibonding state with energy $2E_s + 2\ t_{12}$. The two electrons will enter the lower bonding level and form a covalent bond. This simple picture is very powerful and is normally the way chemists look at the coupling of atoms to each other in chemical complexes. The picture can of course be generalized to different types of bonding orbital and different types of structures. A chemical compound will then normally have the highest occupied level called the HOMO level and lowest unoccupied level called the LUMO level (Fig. 2.13).

Now, if we consider more than two atoms, the effect is similar for a crystalline structure of hydrogen atoms composed of an array of numerous atoms. A general coupling between close atomic orbitals will create new energy levels. The more atoms in the crystalline structure, the more orbital may interact with each other, causing a raise in the number of energy levels. At the limit, when the number of atoms involved in the crystalline structure is high enough, the energy is no longer composed of discrete levels, but forms a quasi-continuum called the energy band. The subject is treated in detail in Chap. 5 .

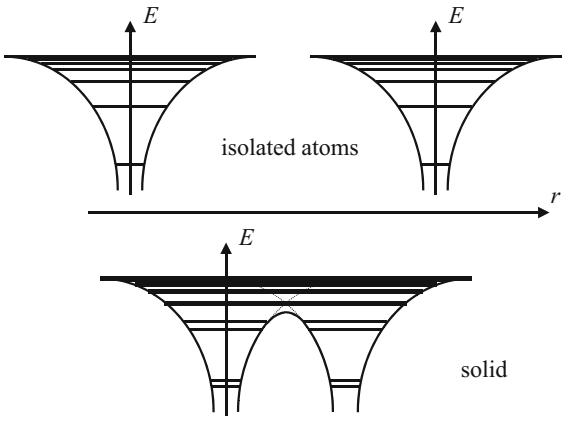


Fig. 2.13 Change in energy spectrum from single atoms to a solid. Each of the discrete energy levels in two isolated atoms split into two separate energy levels when the atoms are bound in a solid. Razeghi FSSE [Razeghi]

In an infinite 3D periodic system of atoms or crystal, the energy band dispersion $E(k)$ due to one orbital per atom for a cubic lattice can be calculated with the tight binding method (see Appendix 2):

$$E\left(\vec{k}\right) = \varepsilon_0 - 2|t|\left(\cos k_x a + \cos k_y b + \cos k_z c\right) \tag{2.1}$$

$$E\left(\vec{k}\right) = \varepsilon_0 - 6|t|\left(1 - \frac{k^2 a^2}{2}\right) \text{when } k \to 0 \tag{2.2}$$

$$\text{where} \frac{1}{m^*} = \frac{1}{\hbar^2}\frac{d^2E}{d^2k_x} \text{ so that} \frac{1}{m^*} = \frac{2\mid t \mid a^2}{\hbar^2} \tag{2.3}$$

where a is the lattice spacing, t is the banding energy, k is the Bloch momentum, m^* is the x-effective mass, and ε_0 is the orbital level. Basically the energy levels of the crystal are labeled by the k-vectors, and the reason why a solution is possible in this neat way is the periodicity. The reader should consult Chap. 5 for a more rigorous discussion. Here the emphasis is on the atomic orbital starting point.

As we can see in Eq. (2.1), the energy $E(k)$ forms a continuum, meaning that for an infinite periodic system, the separation between consecutive energy levels is infinitesimally small. Without electronic correlations, such a solid with the same orbital on every site, with one electron per orbital, would give rise to a metal, because an electric field will easily push a charge away from its starting point without having to surmount a large energy barrier. This is not so for an insulator or semiconductor. A normal size applied field would not be able to move the charge across the energy gap or bandgap. Another important concept is the effective mass m^* defined in Eq. 2.3; it is a measure of the efficiency with which an applied field accelerates an electron if Newton's law were to apply [see Chap. 5 Razeghi or standard quality textbooks in solid-state physics listed below].

Different energy shells in atoms when they couple can be thought to generate their own energy bands. But if in a system with only one orbital per atom, we can still have bandgaps and energy bands. In this case the bandgap of a periodic system arises usually when there is more than one atom per unit cell.

To illustrate this we consider a linear chain constituted of two different atoms alternatively labeled A and B with only one orbital per atom and orbital energies E_A and E_B. Assuming the transfer coupling is called t, then a simple tight binding energy band structure calculation gives us now two energy bands instead of one and given by ($k = k_x$) (Fig. 2.14).

Fig. 2.14 Linear chain of two different atoms

$$2E(k) = (E_A + E_B) \pm \left[(E_A - E_B)^2 + 8t^2(1 + \cos 2ka)\right]^{1/2} \tag{2.4}$$

The lower band can be identified as emanating from the lower orbital energy, say E_A, the upper from B; electron staring in A has to transfer across the higher orbital B to reach the next equivalent site. For the higher band, the result is similar. Assuming one electron per A orbital, and initially none on B, with the Pauli principle, this density fills the lower band completely and leaves the upper band empty at $T = 0$ K. Now we have a semiconductor with a bandgap of $2E_g = [(E_A - E_B)^2 + 16t^2]^{1/2}$.

The argument is easily extended to higher dimensions, and in general, in an infinite lattice, in order to have a bandgap, one has to have more than one atom per unit cell which periodically repeats itself. Thus silicon and germanium have identical atoms, but the Bravais lattice structure implies two atoms per unit cell in an FCC lattice and this gives a semiconductor [see Chap. 5]. If the structure were a simple cubic with one atomic orbital per cell, then there would be no bandgap.

Band structures can be evaluated in a variety of ways depending on the degree of accuracy required. The simplest methods are using the linear combination or atomic orbitals or tight binding method (TBM) or using the free electron representation with atoms acting as scattering centers (see references below) NFE. In complex materials such as transition metals or layered compounds and oxides, one has to use more accurate methods. In so-called ab initio methods, density functional and pseudopotential methods, the objective is to assume as little as possible about the system yet to arrive at accurate pictures of the electronic and phononic structures. The Schrödinger equation SE of the electron (see Chap. 4) is solved in an array of atomic scatterers defined by the lattice structure, whereby only the outermost valence electrons are treated, and the scattering potential of the atom is represented by an effective core due to the non-valence shells and ions. The core is given a radius and spherical magnitude and form often fitted to experiment. The SE is solved using a plane wave basis of say 500 plane waves, and the electron-electron interaction is treated as a self-consistent averaged field using the Hartree or Hartree-Fock theory. It is now more common to solve for the electron density rather than for the wavefunction, using the so-called density functional method DFT (Dreizler 1985).

2.7 Band Structure of Fullerenes (Buckyballs) (Fig. 2.15)

2.8 Band Structure of Carbon Nanotubes (Fig. 2.16)

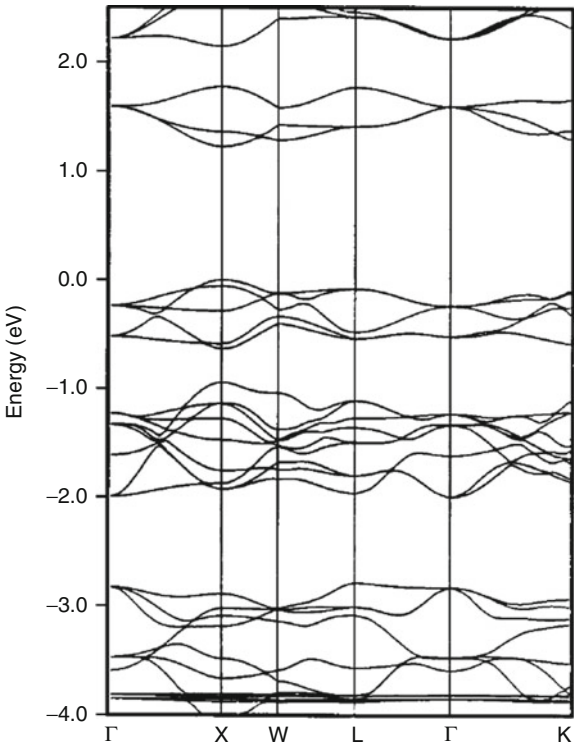


Fig. 2.15 Solid Fcc fullerene shows the band structure of solid undoped fullerenes on an Fcc lattice. This also corresponds to the structure of the superconducting K_3C_{60}. See Gunnarsson (1997). Subbands around the fermi energy for solid C_{60} in the Fm3 structure. The bands at about −0.5 eV are the h_u bands which are occupied in solid C_{60}, and the bands at about 1.5 eV are the t_{1u} bands which become populated in A_nC_{60}. [From Gunnarsson (1997); Erwin and Pederson (1993)]

2.9 Background Needed for Energy Levels and Band Structure

2.9.1 Tight Binding Method

Razeghi M, page 169, Fundamentals of solid state engineering [5]

Peyghambarian N, Koch S and Mysyrowicz A Introduction to semiconductor optics, page 27

2.9.2 Free Electron Method

Razeghi M, Fundamentals of solid state engineering Chapter 4, or other quality solid-state textbooks for more information are by Ashcroft and Mermin

Ziman J (1964) An introduction to Solid State Physics, Cambridge University Press (1964)

Madelung O (1978) Introduction to Solid state theory, Springer Berlin Heidelberg, New York

Ziman J (1964) An introduction to Solid State Physics, Cambridge University Press (1964)

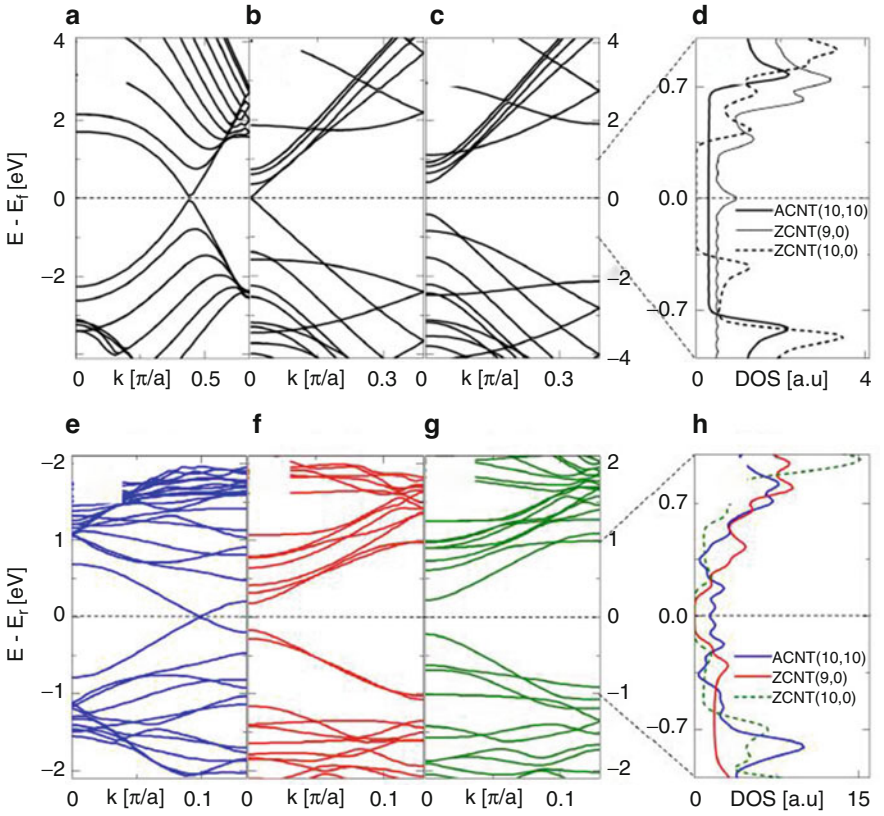


Fig. 2.16 Band structure in pristine CNT (**a**) ACNT (10,10), (**b**) ZNT (9,0), (**c**) ZNT (10,10), (**d**) DOS. Porous CNT (**e**) ACNT (10,10), (**f**) ZCNT (9,0), (**g**) ZCNT (10,10), (**h**) DOS. The energy scales for (**b**, **c**) and (**f**, **g**) cases are the same. The density of states is shown in the right curves (**d**, **h**). *From electronic structure of porous nanocarbons* (Artem Baskin and P Kral, Scientific Reports 1:36 https://doi.org/10.1038/srep00036)

Dreizler M (1985) Density functional theory in Physics NATO ASI series vol. 123

*A more complete description of band structure is given in Chap. 5.

2.10 Summary

In this chapter we reviewed the basic properties of carbon as an atom and looked at the various ways carbon bonds to form chemical complexes and allotropes. This is an ever-evolving field and the material presented is by no means complete. In the next few chapters, we will examine each one of these allotropes in turn and investigate their band structure, electronic properties, and applications. The synthesis and preparation method of each one of these forms is a specialized subject in its own

right, and the interested reader should refer to the numerous and detailed literature on the subject, best done by using Google. In Sect. 2.2 we briefly introduce the reader to the concept of bonding in periodic systems. This gives the so-called Bloch energy band structure $E(k)$ and is very rich in new physical concepts. The Bloch bands also lead us to the concept of holes in the valence bands and the effective mass of the electrons and holes. In this way, we prepare the reader to understand the material to come later in the book, but we expect him to follow up each one of these new concepts in detail with the specialized literature.

2.11 Conclusion: The Future

Carbon atoms exhibit enormous flexibility in the way they bond and form material complexes. Carbon has given rise to astonishing new structures and lead to great dazzling discoveries. One of the greatest breakthroughs came with potassium K_3C_{60} electron-doped fullerenes which exhibited superconductivity up to 40 K [Gunnarsson]. This was an amazing achievement which went above all expectations and which gave H Kroto the Nobel Prize [Kroto] even though the complete understanding of the mechanism has to our knowledge not yet been reached.

Another major success was scored earlier by the work of the Santa Barbara Group under Alan Heeger [Heeger]. They showed that polyacetylene (PA) could be considered to be a quasi one-dimensional polymer and that the material (and many others of this type) could be n and p doped. Su Schrieffer and Heeger [Su et al. 1979] predicted that theoretically, undoped PA should constitute a Peierls semiconductor. In other words a semiconductor where the energy gap forms is a result of a collective relaxation of the backbone into an alternating "short-long" (or long-short) carbon bond structure. Doping then causes a semiconducting to metal transition which is not just purely an electronic transition, but a structural transition as well. Here the alternating bond length changes partially back into the "normally to be expected" (disregarding lattice relaxation) same bond length structure. This novel and collective interplay between lattice and electronic structure also gives rise to very exciting new types of elementary excitations known as "polarons and solitons" [Su et al]. The work on conjugated polymers and applications then eventually gave A Heeger the nobel Prize in the year 2000. The most recent and perhaps one of the most promising discoveries from the point of view of material engineering and applications is the isolation (exfoliation) of graphene sheets from graphite. This truly amazing discovery has made it possible to make 2D pristine monolayer "metallic" materials which have great unprecedented structural stability and therefore technical value. It gave the discoverers Geim and Novoselov the Nobel Prize in 2010 [Geim]. Graphene also exists in stable suspended form and exhibits high mobility (100,000 cm^2/Vs). Graphene has given device technology a new class of field-effect transistors and sensors. It has a zero bandgap not at $k = 0$ (Γ point) but at the so-called Dirac points. Here the zero gap makes linear dispersions, and if one insists on pushing the mass concept to its extreme, we have zero effective mass particles and a square

root magnetic field dependence of magnetic level splittings. Graphene can be made into nanoscale ribbons which introduces lateral quantum confinement and brings back an energy gap which can be adjusted by design. The search for new applications using graphene and related complexes is by far not over. Many groups, including the original discoverers, now in Manchester, are looking for more new science and technical applications. In particular the focus is now on effects related to interlayer electron-electron coupling. Here one is trying to make or observe charge polarization and the drag of the polarization induced in the neighboring layers to form new electronic polarons and bipolarons. There is also still hope that some topology can be found which will eventually yield very high-temperature superconductivity, higher than K_3C_{60}. Organic high-temperature ferromagnetism is still a very sought after target. One of the mysteries of the solid-state physics of carbon is how far one can go with single particle mean field theories. A material with the topology of graphene, for example, would seem to really necessitate a many-body treatment of electronic structures, but apparently this is not the case, and one-body methods work quite well. Whereas in conjugated polymers and molecular structures, electron-phonon and lattice relaxation have been shown to play a serious role in determining energies and structure, the same is not true for electron-electron coupling. Though we know that Coulomb correlations are present and non-negligible, the scope, importance, and deep understanding of correlations are still missing in carbon-based materials. In most current theoretical treatments, correlations can be incorporated into the redefinition of one-body parameters. So a lot more needs to be done in order to come to understand the full potential of "carbon" and related materials. Thus in K_3C_{60} [Gunnarsson], most scientists have been more busy trying to explain away the electron-electron on-site correlation called Hubbard U. This coupling would, for example, act on the fullerene balls and is ~1.5 eV [Gunnarsson]. If correlations could be proven to be instrumental in producing superconductors, as is the case for magnetism, it would open new avenues for materials research. The research could focus, it seems, on looking for more exotic topologies, such as nanocrystalline assemblies quantum dots and crystals and even porous forms. Some of the new imaginative molecular material designs, which chemists are capable of producing, may well eventually give the sought after exciting properties such as high-temperature superconductivity and lightweight magnetism, including both ferro- and diamagnetism. Luminescent carbon nanodots have already been delivered [Baker], thanks to the discovery regarding the effect of passivation. This field still has a lot of potential since the complete mechanisms are still not understood, and wavelength control may be possible. The search is on and is exciting. But electronic structure is only one aspect, and carbon allotropes, because of this unusual structural mechanical strength, are proving extremely valuable in fields such as civil engineering, aircraft, and car manufacturing. Not all facets and combinations of properties (e.g., solar cells, thermal and sound conductivity, and insulation) have been investigated, the potential is enormous, and the development of these fields is of great value to the manufacturing building and automotive and transport industries.

References for Conclusions

Geim AK (2011) Nobel lecture, *Random walk to graphene* Rev. Mod Phys 83.

Su W, Schrieffer JR, Heeger AJ (1979) Solitons In Polyatcetylene PRL 42:1698.

Kroto, HW et al (1985) *C_{60}: Buckminsterfullerene. Nature* 318(6042):162–163. Bibcode:1985Natur.318..162 K. https://doi.org/10.1038/318162a0.

Gunnarsson O (1997) *Rev. of modern physics* 69:575. *Superconductivty in fullerides.*

From A J Heeger Adv Mater., 1, 2013 Wiley –VCH online library "25th Anniversary article: bulk heterojunction solar cells: understanding the mechanism of operation".

Luminescent Carbon Nanodots: Emergent Nanolights S Baker and Gary Baker Angew. Chem Int Ed Nanotechnology 49:6726 (2010).

Problems

Q1. Illustrate the various bonding configurations that carbon can adopt and give examples of materials for each case. Where do you think organic carbon technology can become superior to inorganic technology?

Q2. Explain how sp^3 and sp^2 hybridizations work? How does hybridization work in Si, Ge and in III–V compounds? In an ab initio band structure calculation, the concept of hybridization does not arise; explain the difference.

Q3. Explain how can we calculate the bonding energy between different atoms given the atomic orbital energies of each orbitals.

Q4. What is Hund's rule coupling?

Q5. Calculate the dispersion equation in Appendix 2 example for a 3 dimensional crystal. This equation is used in this chapter in (2.4).

References

Dreizler M (1985) Density functional theory in Physics NATO ASI series vol. 123

Forro L, Mihaly L (2001) Electronic properties of doped fullerenes. Rep Prog Phys 64:649

Gunnarsson O (1997) Superconductivty in fullerides. Rev Mod Phys 69:575

Erwin S, Pederson M (1993) K_3C_{60} Bandstructure"Electronic structure of carbon nanotubes systems measured with scanning tunneling microscopy". PRB 47:14657

Further Reading

Ashcroft NW, Mermin ND (1976) Solid state physics. Holt, Rinehart and Winston, New York. ISBN 0030839939, 9780030839931

Audi G, Wapstra AH, Thibault C, Blachot J Bersillon O (2003) The NUBASE evaluation of nuclear and decay properties Bruckner, R Advanced organic chemistry ISBN 978021381103

Cottrell RT (1958) The strengths of chemical bonds, 2nd edn. Butterworths, London

Darwent B (1970) National standard reference data series, National Bureau of Standards, No. 31, Washington, DC

Demarchi D, Tagliaferro A (n.d.) Carbon for sensing devices details. Springer ISBN 978-3-319-08648-4

de Laeter R, Böhlke JK, De Bièvre P, Hidaka H, Peiser HS, Falkowski P, Scholes RJ, Boyle E, Canadell J, Canfield D, Elser Gruber N, Hibbard K et al (2000) The global carbon cycle: a test of our knowledge of earth as a system. Pure and Appl. Chem, Springer, Vol 75, 8683 2003 IUPAC

Madelung O (1978) Introduction to solid state theory. Springer, Berlin Heidelberg/New York

Parr RG, Yang W Density functional theory of atoms and molecules. Oxford University press, Oxford

Peyghambarian N, Koch S, Mysyrowicz A (1993) Introduction to semiconductor optics. Prentice Hall, Englewood Cliffs New Jersey, Prentice Hall Series in Solid State Electronics

Razeghi M (2009) Fundamentals of solid state engineering 3rd Ed Springer press 2009 FSSE or this book

Rosman KJR, Taylor PDP (2003) Atomic weights of the elements. Review 2000 (IUPAC Technical Report), Pure and Applied Chemistry

Wieser ME (2006) Atomic weights of the elements 2005 (IUPAC Technical Report), Pure and Applied Chemistry

Ziman J (1964) An introduction to solid state physics. Cambridge University press, Cambridge

Crystalline Properties of Solids

3

3.1 Introduction

This chapter gives a brief introduction to crystallography, which is the science that studies the structure and properties of the crystalline state of matter. We will first discuss the arrangements of atoms in various solids, distinguishing between single crystals and other forms of solids. We will then describe the properties that result from the periodicity in crystal lattices. A few important crystallography terms most often found in solid state devices will be defined and illustrated in crystals having basic structures. These definitions will then allow us to refer to certain planes and directions within a lattice of arbitrary structure.

Investigations of the crystalline state have a long history. Johannes Kepler (*Strena Seu de Nive Sexangula*, 1611) speculated on the question as to why snowflakes always have six corners, never five or seven (Fig. 3.1). It was the first treatise on geometrical crystallography. He showed how the close-packing of spheres gave rise to a six-corner pattern. Next, Robert Hooke (*Micrographia*, 1665) and Rene Just Haüy (*Essai d'une théorie sur la structure des cristaux*, 1784) used close-packing arguments in order to explain the shapes of a number of crystals. These works laid the foundation of the mathematical theory of crystal structure. It is only recently, thanks to x-ray and electron diffraction techniques, that it has been realized that most materials, including biological objects, are crystalline or partly so (Fig. 3.2).

All elements from the periodic table and their compounds, be they gas, liquid, or solid, are composed of atoms, ions, or molecules. Matter is discontinuous. However, since the sizes of the atoms, ions, and molecules lie in the 1 Å (10^{-10} m or 10^{-8} m) region, matter appears continuous to us. The different states of matter may be distinguished by their tendency to retain a characteristic volume and shape. A gas adopts both the volume and shape of its container, a liquid has a constant volume but adopts the shape of its container, while a solid retains both its shape and volume independently of its container. This is illustrated in Fig. 3.3. The natural forms of each element in the periodic table are given in Fig. A1 in Appendix A.3.

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_3

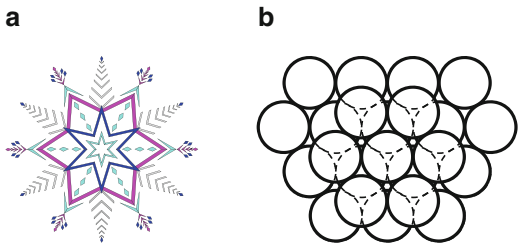


Fig. 3.1 (**a**) Snowflake crystal and (**b**) the close-packing of spheres which gives rise to a six-corner pattern. The close-packing of spheres can be thought as the way to most efficiently stack identical spheres

Gases Molecules or atoms in a gas move rapidly through space and thus have a high kinetic energy. The attractive forces between molecules are comparatively weak and the energy of attraction is negligible in comparison to the kinetic energy.

Liquids As the temperature of a gas is lowered, the kinetic energies of the molecules or atoms decrease. When the boiling point (Fig. A.3 in Appendix A.3) is reached, the kinetic energy will be equal to the energy of attraction among the molecules or atoms. Further cooling thus converts the gas into a liquid. The attractive forces cause the molecules to "touch" one another. They do not, however, maintain fixed positions. The molecules change positions continuously. Small regions of order may indeed be found (local ordering), but if a large enough volume is considered, it will also be seen that liquids give a statistically homogeneous arrangement of molecules and therefore also have isotropic physical properties, i.e., equivalent in all directions. Some special types of liquids that consist of long molecules may reveal anisotropic properties (e.g., liquid crystals).

Solids When the temperature falls below the freezing point, the kinetic energy becomes so small that the molecules become permanently attached to one another. A three-dimensional framework of net attractive interaction forms among the molecules and the array becomes solid. The movement of molecules or atoms in the solid now consists only of vibrations about some fixed positions. A result of these permanent interactions is that the molecules or atoms have become ordered to some extent. The distribution of molecules is no longer statistical but is almost or fully periodically homogeneous, and periodic distribution in three dimensions may be formed.

The distribution of molecules or atoms, when a liquid or a gas cools to the solid state, determines the type of solid. Depending on how the solid is formed, a compound can exist in any of the three forms in Fig. 3.3. The ordered crystalline phase is the stable state with the lowest internal energy (absolute thermal equilibrium). The solid in this state is called the single crystal form. It has an exact periodic arrangement of its building blocks (atoms or molecules).

Sometimes the external conditions at a time of solidification (temperature, pressure, cooling rate) are such that the resulting materials have a periodic arrangement of atoms which is interrupted randomly along two-dimensional sections that can intersect, thus dividing a given volume of a solid into a number of smaller single crystalline regions or grains. The size of these grains can be as small as several

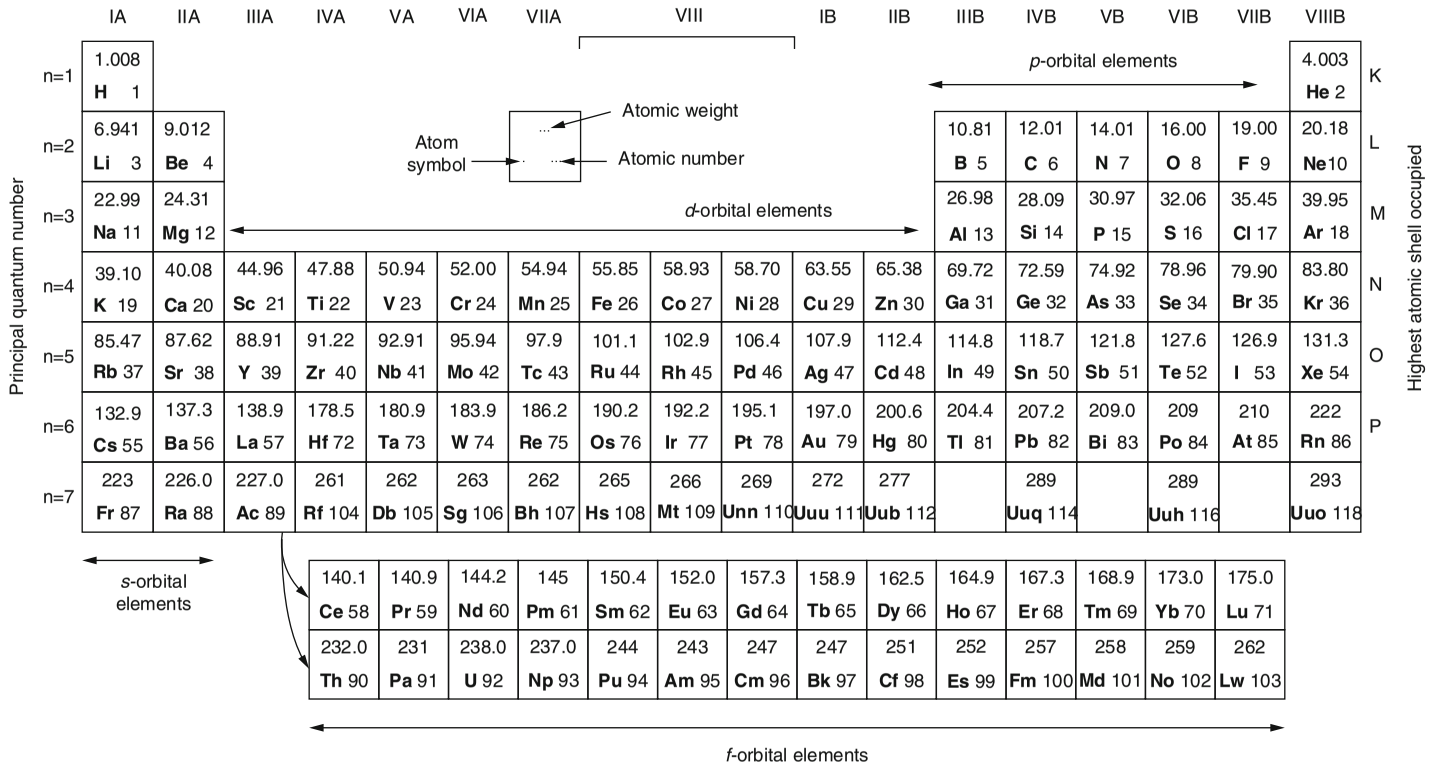


Fig. 3.2 Periodic table of elements. For each element, its symbol, atomic number, and atomic weight are shown

Fig. 3.3 Illustration of the physical states of water: (**a**) gas also known as water vapor, (**b**) liquid or common water, (**c**) solid also known as snow or ice

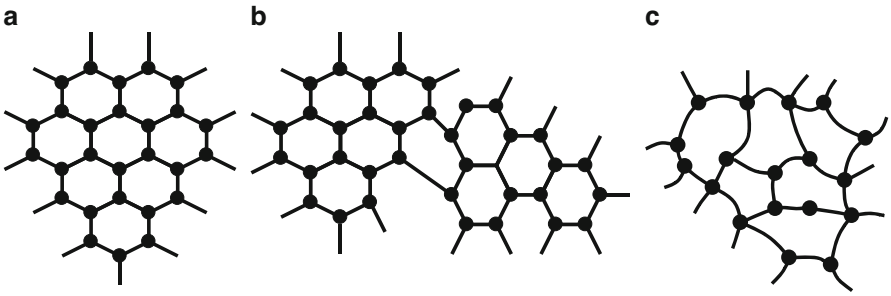


Fig. 3.4 Arrangement of atoms: (**a**) a single crystalline, (**b**) a polycrystalline, and (**c**) an amorphous material

atomic spacings. Materials in this state do not have the lowest possible internal energy but are stable, being in so-named local thermal equilibrium. These are polycrystalline materials.

There exist, however, solid materials which never reach their equilibrium condition, e.g., glasses or amorphous materials. Molten glass is very viscous and its constituent atoms cannot come into a periodic order (reach equilibrium condition) rapidly enough as the mass cools. Glasses have a higher energy content than the corresponding crystals and can be considered as a frozen, viscous liquid. There is no periodicity in the arrangement of atoms (the periodicity is of the same size as the atomic spacing) in the amorphous material. Amorphous solids or glass have the same properties in all directions (they are isotropic), like gases and liquids.

Therefore, the elements and their compounds in a solid state, including silicon, can be classified as single crystalline, polycrystalline, or amorphous materials. The differences among these classes of solids are shown schematically for a two-dimensional arrangement of atoms in Fig. 3.4.

3.2 Crystal Lattices and the Seven Crystal Systems

Now we are going to focus our discussion on crystals and their structures. A crystal can be defined as a solid consisting of a pattern that repeats itself periodically in all three dimensions. This pattern can consist of a single atom, group of atoms, or other

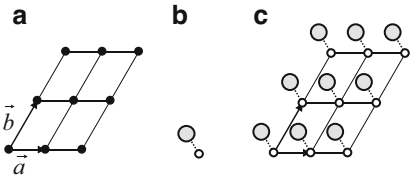


Fig. 3.5 Example of (**a**) two-dimensional lattice, (**b**) pattern, and (**c**) two-dimensional crystal illustrating a pattern associated with each lattice point

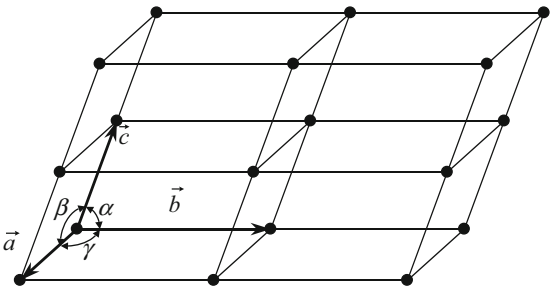


Fig. 3.6 Example of a three-dimensional lattice, with translation vectors and the angles between two vectors. By taking the origin at one lattice point, the position of any lattice point can be determined by a vector which is the sum of integer numbers of translation vectors

compounds. The periodic arrangement of such patterns in a crystal is represented by a lattice. A lattice is a mathematical object which consists of a periodic arrangement of points in all directions of space. One pattern is located at each lattice point. An example of a two-dimensional lattice is shown in Fig 3.5a. With the pattern shown in Fig. 3.5b, one can obtain the two-dimensional crystal in Fig. 3.5c which shows that a pattern is associated with each lattice point.

A lattice can be represented by a set of translation vectors as shown in the two-dimensional (vectors $\vec{a}$, $\vec{b}$) and three-dimensional lattices (vectors $\vec{a}$, $\vec{b}$, $\vec{c}$) in Fig. 3.6a, c, respectively. The lattice is invariant after translations through any of these vectors or any sum of an integer number of these vectors. When an origin point is chosen at a lattice point, the position of all the lattice points can be determined by a vector which is the sum of integer numbers of translation vectors. In other words, any lattice point can generally be represented by a vector $\vec{R}$ such that:

$$\begin{aligned} &\vec{R} = n_1\,\vec{a} + n_2\,\vec{b} + n_3\,\vec{c}\,, \\ &n_{1,2,3} = 0, \pm 1, \pm 2, \ldots \end{aligned} \tag{3.1}$$

where $\vec{a}$, $\vec{b}$, $\vec{c}$ are the chosen translation vectors and the numerical coefficients are integers.

All possible lattices can be grouped in the seven crystal systems shown in Table 3.1, depending on the orientations and lengths of the translation vectors. No crystal may have a structure other than one of those in the seven classes shown in Table 3.1.

Table 3.1 The seven crystal systems

Crystal systems	Axial lengths and angles
Cubic	Three equal axes at right angles $a = b = c$, $\alpha = \beta = \gamma = 90°$
Tetragonal	Three axes at right angles, two equal $a = b \neq c$, $\alpha = \beta = \gamma = 90°$
Orthorhombic	Three unequal axes at right angles $a \neq b \neq c$, $\alpha = \beta = \gamma = 90°$
Trigonal	Three equal axes, equally inclined $a = b = c$, $\alpha = \beta = \gamma = 90°$
Hexagonal	Two equal coplanar axes at 120°, third axis at right angles $a = b \neq c$, $\alpha = \beta = 90°$, $\gamma = 120°$
Monoclinic	Three unequal axes, one pair not at right angles $a \neq b \neq c$, $\alpha = \gamma = 90° \neq \beta$
Triclinic	Three unequal axes, unequally inclined and none at right angles $a \neq b \neq c$, $\alpha \neq \beta \neq \gamma \neq 90°$

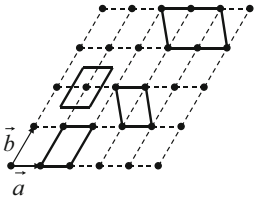


Fig. 3.7 Three examples of possible unit cells for a two-dimensional lattice The unit cells are delimited in solid lines. The same principle can be applied for the choice of a unit cell in three dimensions

A few examples of cubic crystals include Al, Cu, Pb, Fe, NaCl, CsCl, C (diamond form), Si, and GaAs; tetragonal crystals include In, Sn, and TiO_2; orthorhombic crystals include S, I, and U; monoclinic crystals include Se and P; triclinic crystals include $KCrO_2$; trigonal crystals include As, B, and Bi; and hexagonal crystals include Cd, Mg, Zn, and C (graphite form) (Fig. 3.6).

3.3 The Unit Cell Concept

A lattice can be regarded as a periodic arrangement of identical cells offset by the translation vectors mentioned in the previous section. These cells fill the entire space with no void. Such a cell is called a unit cell.

Since there are many different ways of choosing the translation vectors, the choice of a unit cell is not unique and all the unit cells do not have to have the same volume (area). Figure 3.7 shows several examples of unit cells for a two-dimensional lattice. The same principle can be applied when choosing a unit cell for a three-dimensional lattice.

The unit cell which has the smallest volume is called the primitive unit cell. A primitive unit cell is such that every lattice point of the lattice, without exception, can be represented by a vector such as the one in Fig. 3.7. An example of primitive unit cell in a three-dimensional lattice is shown in Fig. 3.6. The vectors defining the unit cell, $\vec{a}$, $\vec{b}$, $\vec{c}$, are basis lattice vectors of the primitive unit cell.

The choice of a primitive unit cell is not unique either, but all possible primitive unit cells are identical in their properties: they have the same volume, and each

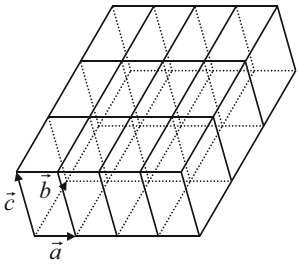


Fig. 3.8 Three-dimensional lattice and a corresponding primitive unit cell defined by the three basis vectors $\vec{a}$, $\vec{b}$, $\vec{c}$

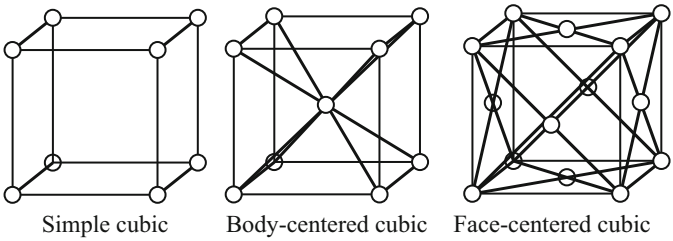


Fig. 3.9 Three-dimensional unit cells: simple cubic (left), body-centered cubic (bcc) (middle), and face-centered cubic (fcc) (right)

contains only one lattice point. The volume of a primitive unit cell is found from vector algebra:

$$V = \left| \vec{a} \cdot \left(\vec{b} \times \vec{c} \right) \right| \tag{3.2}$$

The number of primitive unit cells in a crystal, N, is equal to the number of atoms of a particular type, with a particular position in the crystal, and is independent of the choice of the primitive unit cell:

$$\text{Primitive unit cell volume} = \frac{\text{Crystal volume}}{N}$$

A primitive unit cell is in many cases characterized by non-orthogonal lattice vectors (as in Fig. 3.8). As one likes to visualize the geometry in orthogonal coordinates, a conventional unit cell (but not necessarily a primitive unit cell) is often used. In most semiconductor crystals, such a unit cell is chosen to be a cube, whereas the primitive cell is a parallelepiped and is more convenient to use due to its more simple geometrical shape.

A conventional unit cell may contain more than one lattice point. To illustrate how to count the number of lattice points in a given unit cell, we will use Fig. 3.9 which depicts different cubic unit cells.

In our notations n_i is the number of points in the interior, n_f is the number of points on faces (each n_f is shared by two cells), and n_c is the number of points on corners

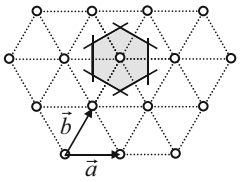


Fig. 3.10 Two-dimensional Wigner-Seitz cell and its construction method: select a lattice point, draw lines from a given lattice point to all nearby points, bisect these lines with orthogonal planes, and construct the smallest polyhedron that contains the first selected lattice

(each n_c point is shared by eight corners). For example, the number of atoms per unit cell in the fcc lattice (Fig. 3.9c) ($n_i = 0$, $n_f = 6$, and $n_c = 8$) is:

$$n_u = n_i + \frac{n_f}{2} + \frac{n_c}{8} = 4 \text{ atoms/unit cell} \tag{3.3}$$

3.4 The Wigner-Seitz Cell

The primitive unit cell that exhibits the full symmetry of the lattice is called Wigner-Seitz cell. As it is shown in Fig. 3.10, the Wigner-Seitz cell is formed by (1) drawing lines from a given Bravais lattice point to all nearby lattice points, (2) bisecting these lines with orthogonal planes, and (3) constructing the smallest polyhedron that contains the selected point. This construction has been conveniently shown in two dimensions but can be continued in the same way in three dimensions. Because of the method of construction, the Wigner-Seitz cell translated by all the lattice vectors will exactly cover the entire lattice.

3.5 Bravais Lattices

Because a three-dimensional lattice is constituted of unit cells which are translated from one another in all directions to fill up the entire space, there exist only 14 different such lattices. They are illustrated in Fig. 3.11 and each is called a Bravais lattice after the name of Bravais (1848).

In the same manner, as no crystal may have a structure other than one of those in the seven classes shown in Fig. 3.11, no crystal can have a lattice other than one of those 14 Bravais lattices.

3.6 Point Groups

Because of their periodic nature, crystal structures are brought into self- coincidence under a number of symmetry operations. The simplest and most obvious symmetry operation is translation. Such an operation does not leave any point of the lattice

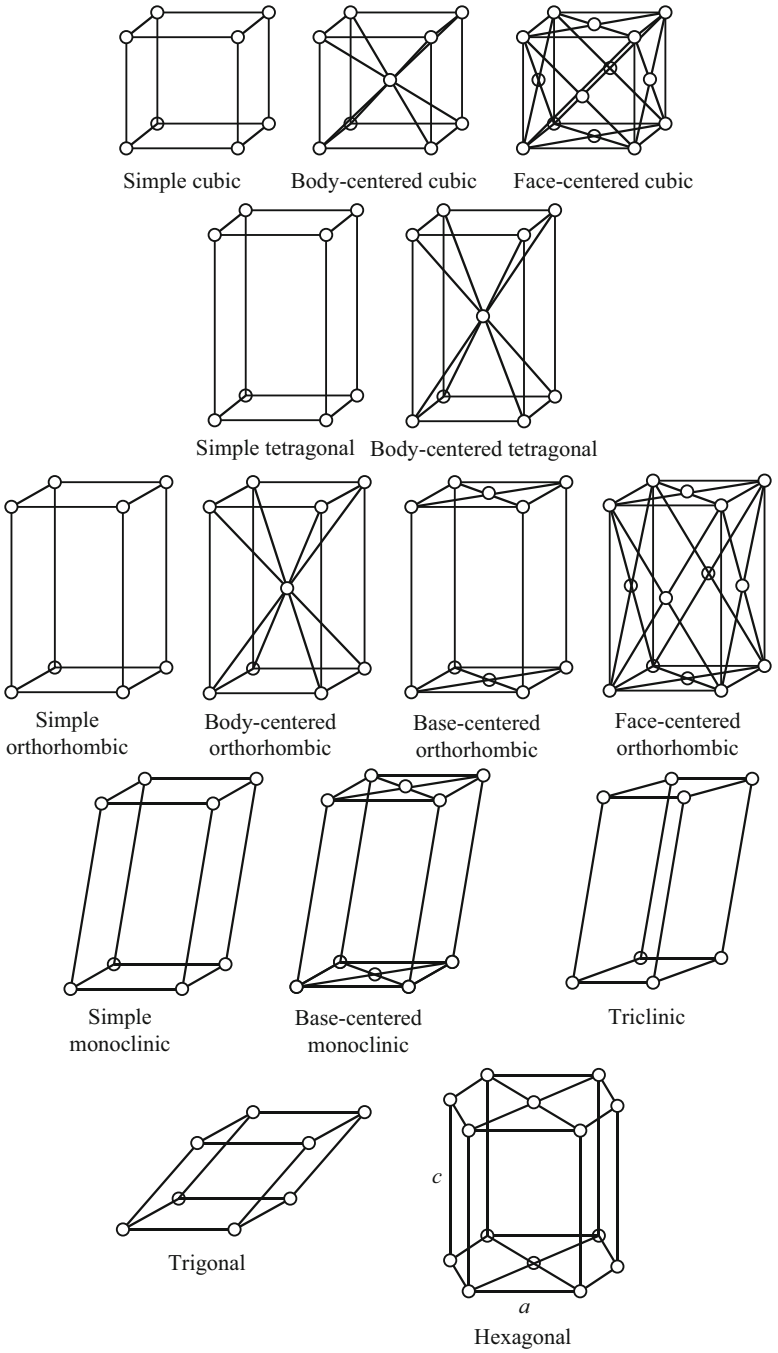


Fig. 3.11 The 14 Bravais lattices, illustrating all the possible three-dimensional crystal lattices

invariant. There exists another type of symmetry operation, called point symmetry, which leaves a point in the structure invariant. All the point symmetry operations can be classified into mathematical groups called point groups, which will be reviewed in this section.

The interested reader is referred to mathematics texts on group theory for a complete understanding of the properties of mathematical groups. For the scope of the discussion here, one should simply know that a mathematical group is a collection of elements which can be combined with one another and such that the result of any such combination is also an element of the group. A group contains a neutral element such that any group element combined with it remains unchanged. For each element of a group, there also exists an inverse element in the group such that their combination is the neutral element.

3.6.1 C_s Group (Plane Reflection)

A plane reflection acts such that each point in the crystal is mirrored on the other side of the plane as shown in Fig. 3.12. The plane of reflection is usually denoted by σ. When applying the plane reflection twice, i.e., σ^2, we obtain the identity which means that no symmetry operation is performed. The reflection and the identity form the point group which is denoted C_s and which contains only these two symmetry operations (Fig. 3.13).

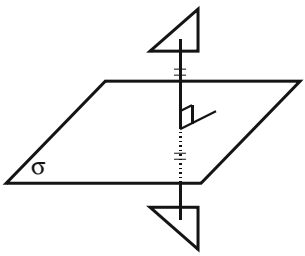


Fig. 3.12 Illustration of a plane reflection. The triangular object and its reflected image are mirror images of each other

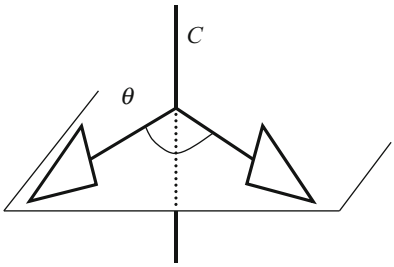


Fig. 3.13 Illustration of rotation symmetry. The triangular object and its image are separated by an angle equal to θ

3.6.2 C_n Groups (Rotation)

A rotation about an axis and through an angle θ (n is an integer) is such that any point and its image are located in a plane perpendicular to the rotation axis and the in-plane angle that they form is equal to θ, as shown in Fig. 3.14. In crystallography, the angle of rotation cannot be arbitrary but can only take the following fractions of 2π: $\theta = \frac{2\pi}{1}$, $\frac{2\pi}{2}$, $\frac{2\pi}{3}$, $\frac{2\pi}{4}$, $\frac{2\pi}{6}$.

It is thus common to denote as C_n a rotation through an angle $\frac{2\pi}{n}$ where n is an integer equal to 1, 2, 3, 4, or 6. The identity or unit element corresponds to $n = 1$, i.e., C_1. For a given axis of rotation and integer n, a rotation operation can be repeated, and this actually leads to n rotation operations about the same axis, corresponding to the n allowed angles of rotation: $1 \times \frac{2\pi}{n}$, $2 \times \frac{2\pi}{n}$, ..., $(n-1) \times \frac{2\pi}{n}$, and $n \times \frac{2\pi}{n}$. These n rotation operations, which include the identity, form a group also denoted C_n.

One says that the C_n group consists of n-fold symmetry rotations, where n can be equal to 1, 2, 3, 4, or 6. Figure 3.14 depicts the perspective view of the crystal bodies with symmetries C_1, C_2, C_3, C_4, C_6. The rotations are done so that the elbow pattern coincides with itself. It is also common to represent these symmetry groups with the rotation axis perpendicular to the plane of the figure, as shown in Fig. 3.15.

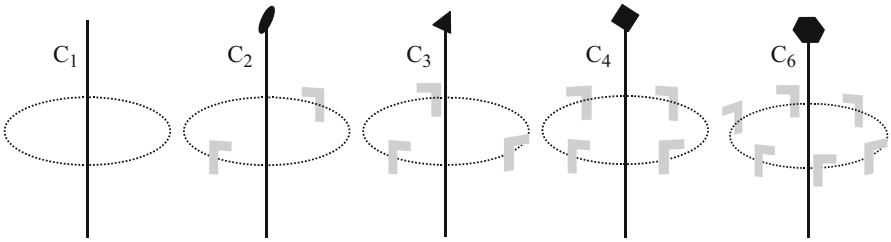


Fig. 3.14 Crystal bodies with symmetries C_1, C_2, C_3, C_4, and C_6. The elbow patterns are brought into self-coincidence after a rotation around the axis shown and through an angle equal to $2\pi/n$ where $n = 1, 2, 3, 4$, or 6

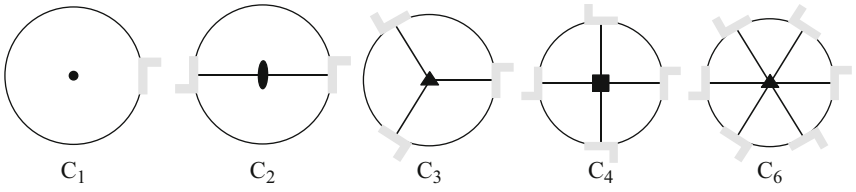


Fig. 3.15 Crystal bodies with symmetries C_1, C_2, C_3, C_4, and C_6 with the rotation axes perpendicular to the plane of the figure

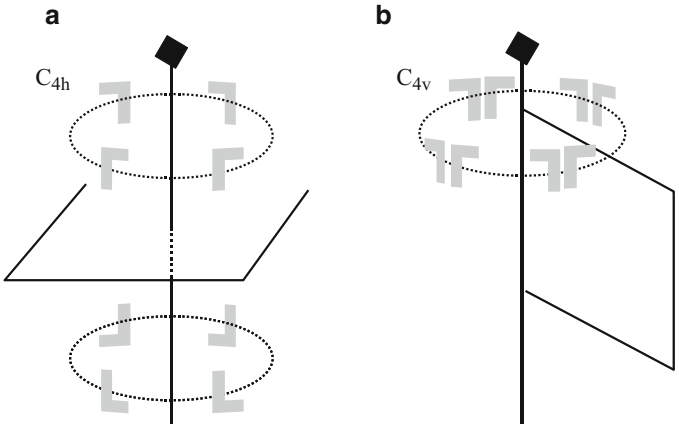


Fig. 3.16 Crystal bodies with symmetries (**a**) C_{4h} where the reflection plane is perpendicular to the rotation axis and (**b**) C_{4v} where the reflection plane passes through the rotation axis

3.6.3 C_{nh} and C_{nv} Groups

When combining a rotation of the C_n group and a reflection plane σ, the axis of rotation is usually chosen vertical. The reflection plane can either be perpendicular to the axis and then be denoted σ_h (horizontal) or pass through this axis and then be denoted σ_v (vertical). All the possible combinations of such symmetry operations give rise to two types of point groups: the C_{nh} and the C_{nv} groups.

The C_{nh} groups contain an n-fold rotation axis C_n and a plane σ_h perpendicular to it. (a) Shows the bodies with a symmetry C_{4h}. The number of elements in a C_{nh} group is $2n$.

The C_{nv} groups contain an n-fold axis C_n and a plane σ_v passing through the rotation axis. Figure 3.16b shows the bodies with a symmetry C_{4v}. The number of elements is $2n$ too.

3.6.4 D_n Groups

When combining a rotation of the C_n group and a C_2 rotation with an axis perpendicular to the first rotation axis, this gives rise to a total of n C_2 rotation axes. All the possible combinations of such symmetry operations give rise to the point groups denoted D_n. The number of elements in this point group is $2n$. For example, the symmetry operations in D_4 are illustrated in Fig. 3.17.

3.6.5 D_{nh} and D_{nd} Groups

When combining an element of the C_{nh} group and a C_2 rotation which has an axis perpendicular to the C_n axis, this gives also rise to a total of n C_2 rotation axes. All

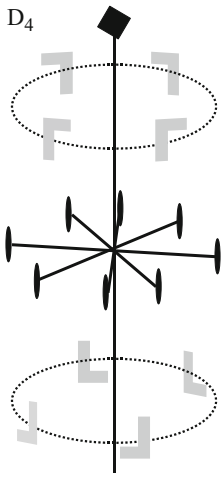


Fig. 3.17 Crystal bodies with symmetry D_4. In addition to the C_4 axis, there are four C_2 axes of rotation perpendicular to the C_n axis

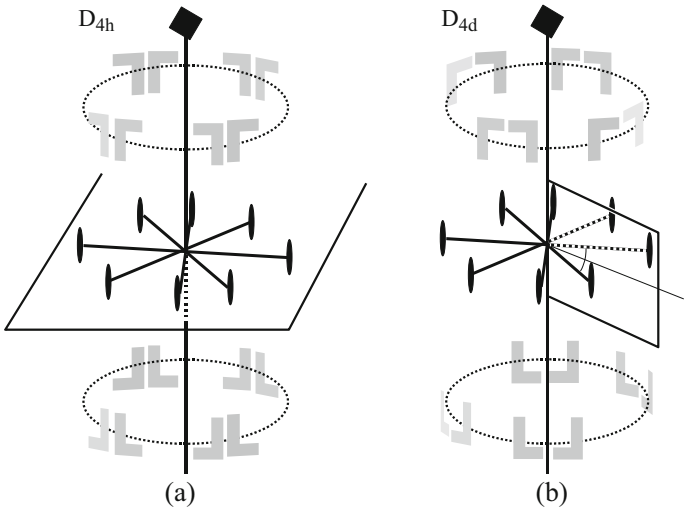


Fig. 3.18 Bodies with symmetries (**a**) D_{4h} and (**b**) D_{4d}

the possible combinations of such symmetry operations lead to the point group denoted D_{nh}. This point group can also be viewed as the result of combining an element of the D_n group and a σ_h (horizontal) reflection plane. This group can also be viewed as the result of combining an element of the D_n group and n σ_v (vertical) reflection planes which pass through both the C_n and the n C_2 axes.

The number of elements in the D_{nh} point group is $4n$, as it includes the $2n$ elements of the D_n group, and all these $2n$ elements combined with a plane reflection σ_h. For example, the symmetry operations in D_{4h} are illustrated in Fig. 3.18a.

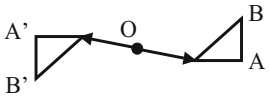


Fig. 3.19 Illustration of inversion symmetry. Any point of the triangular object and its image are such that the inversion center is at the middle of these two points

Now, when combining an element of the C_{nv} group and a C_2 rotation which has an axis perpendicular to the C_n axis and which is such that the σ_v (vertical) reflection planes bisect two adjacent C_2 axes, this leads to the point group denoted D_{nd}. This point group can also be viewed as the result of combining an element of the D_n group and n σ_v (vertical) reflection planes which bisect the C_2 axes.

The number of elements in the D_{nd} point group is $4n$ as well. For example, the symmetry operations in D_{4d} are illustrated in 3.18b.

3.6.6 C_i Group

An inversion symmetry operation involves a center of symmetry (e.g., O) which is at the middle of a segment formed by any point (e.g., A) and its image through inversion symmetry (e.g., A′), as shown in Fig. 3.19.

When applying an inversion symmetry twice, we obtain the identity which means that no symmetry operation is performed. The inversion and the identity form the point group which is denoted C_i and which contains only these two symmetry operations.

3.6.7 C_{3i} and S_4 Groups

When combining an element of the C_n group and an inversion center located on the axis of rotation, the symmetry operations get more complicated. If we consider the C_1 group (identity), we obtain the inversion symmetry group C_i. In the case of C_2 group, we get the plane reflection group C_s. And if we consider the C_6 group, we actually obtain the C_{3h} point group.

When we combine *independently* elements from the C_4 group and the inversion center, we get the C_{4h} point group. However, there is a subgroup of the C_{4h} point group which can be constructed by considering a new symmetry operation, the roto-inversion, which consists of a C_4 rotation immediately followed by an inversion through a center on the rotation axis. It is important to realize that the roto-inversion is a single symmetry operation, i.e., the rotation is not independent of the inversion. The subgroup is made by combining roto-inversion operation, is denoted S_4, and is illustrated in Fig. 3.20. Its number of elements is 4.

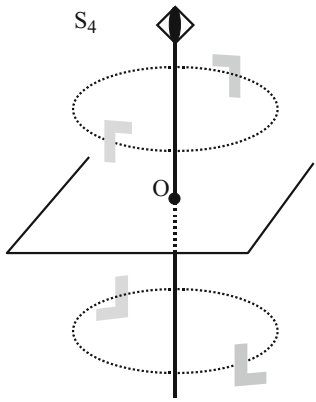


Fig. 3.20 Bodies with symmetry S_4

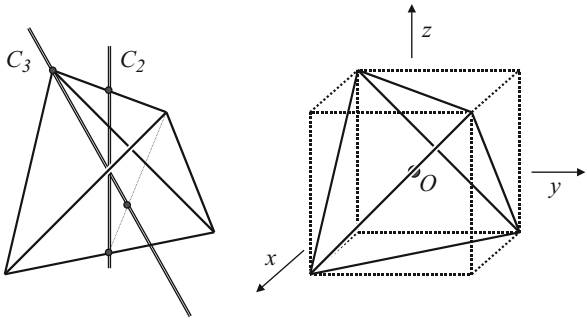


Fig. 3.21 Axes of rotation for the T group, including four C_3 and three C_2 axes. The orientation of the tetrahedron with respect to the cubic coordinate axes is shown on the right

A similar point group is obtained when considering roto-inversions from the C_3 group. The new point group is denoted C_{3i}.

3.6.8 *T* Group

The tetrahedron axes group T is illustrated in Fig. 3.21. It contains some of the symmetry operations which bring a regular tetrahedron into self-coincidence. The tetrahedron and its orientation with respect to the cubic coordinate axes are also shown.

The number of elements is 12, which includes:

- Rotations through an angle $\frac{2\pi}{3}$ or $\frac{4\pi}{3}$, about the four C_3 axes which are the body diagonals of a cube (yielding at total of eight elements)
- Rotations through an angle π, about the three C_2 axes $(\vec{x},\vec{y},\vec{z})$ passing through the centers of opposite faces (three elements)
- The identity (one element)

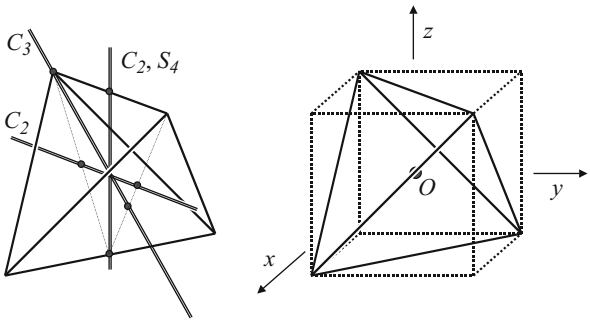


Fig. 3.22 Axes of rotation for T_d group, including four C_3, three C_2 axes passing through the center of opposite faces, three S_4 axes, and six C_2 axes passing through the centers of diagonally opposite sides

3.6.9 T_d Group

The T_d point group contains all the symmetry elements of a regular tetrahedron. Basically, it includes all the symmetry operations of the T group in addition to an inversion center at the center of the tetrahedron (Fig. 3.22).

The number of elements is 24, which includes:

- Rotations through an angle $\frac{2\pi}{3}$ or $\frac{4\pi}{3}$, about the four C_3 axes which are the body diagonals of a cube (yielding at total of eight elements)
- Rotations through an angle π, about the three C_2 axes $(\vec{x},\vec{y},\vec{z})$ passing through the centers of opposite faces (three elements)
- Rotations through an angle $\frac{\pi}{2}$ or $\frac{3\pi}{2}$ (S_4), about the three axes $(\vec{x},\vec{y},\vec{z})$ passing through the centers of opposite faces, followed by an inversion through the center point O of a cube, (six elements)
- Rotations through an angle π, about the six C_2 axes passing through the centers of diagonally opposite sides (in diagonal planes of a cube), followed by an inversion through the center point O (six elements)
- Finally, the identity (one element)

3.6.10 O Group

The cubic axes group O consists of rotations about all the symmetry axes of a cube. The number of elements is 24, which includes:

- Rotations through the angles $\frac{2\pi}{4}$, $\frac{4\pi}{4}$, or $\frac{6\pi}{4}$, about the three C_4 axes passing through the centers of opposite faces (yielding a total of nine elements)
- Rotations through the angles $\frac{2\pi}{3}$ or $\frac{4\pi}{3}$, about the four C_3 axes passing through the opposite vertices (eight elements)
- Rotations through an angle π, about the six C_2 axes passing through the midpoints of opposite edges (six elements)
- Finally, the identity (one element)

3.6.11 O_h Group

The O_h group includes the full symmetry of a cube in addition to an inversion symmetry. The number of elements is 48, which includes:

- All the symmetry operations of the O group (24 elements)
- And all the symmetry operations of the O group combined with an inversion through the body-centered point of a cube (24 elements).

3.6.12 List of Crystallographic Point Groups

The point groups previously reviewed are constructed by considering all the possible combinations of basic symmetry operations (plane reflections and rotations) discussed in subsections 3.6.1 to 3.6.11. By doing so, one would find that there exist only 32 crystallographic point groups. Crystallographers normally use two kinds of notations for these point symmetry groups. Table 3.2 shows the correspondence between two widely used notations.

3.7 Space Groups

The other type of symmetry in crystal structures, (translation symmetry), reflects the self-coincidence of the structure after the displacements through arbitrary lattice vectors ($\vec{R}$).

These symmetry operations are independent of the point symmetry operations as they do not leave a point invariant (except for the identity). The combination of translation symmetry and point symmetry elements gives rise to new symmetry operations which also bring the crystal structure into self-coincidence. An example of such new operation is a glide plane by which the structure is reflected through a reflection plane and then translated by a vector parallel to the plane.

With these new symmetry operations, a larger symmetry operation group is formed, called space group. There are only 230 possible three-dimensional crystallographic space groups which are conventionally labeled with a number from No. 1 to No. 230.

3.8 Directions and Planes in Crystals: Miller Indices

In order to establish the proper mathematical description of a lattice, we have to identify the directions and planes in a lattice. This is done in a crystal using Miller indices (*hkl*). We introduce Miller indices by considering the example shown in Fig. 3.23.

Table 3.2 List of the 32 crystallographic point groups

Crystal system	Schoenflies symbol	Hermann-Mauguin symbol
Triclinic	C_1	*1*
	C_i	$\bar{1}$
Monoclinic	C_2	*2*
	C_s	*m*
	C_{2h}	*2/m*
Orthorhombic	D_2	*222*
	C_{2v}	*mm2*
	D_{2h}	*mmm*
Tetragonal	C_4	*4*
	S_4	$\bar{4}$
	C_{4h}	*4/m*
	D_4	*422*
	C_{4v}	*4 mm*
	D_{2d}	$\bar{4}2m$
	D_{4h}	*4/mmm*
Cubic	T	*23*
	T_h	*m3*
	O	*432*
	T_d	$\bar{4}3m$
	O_h	*m3m*
Trigonal	C_3	*3*
	C_{3i}	$\bar{3}$
	D_3	*32*
	C_{3v}	*3m*
	D_{3d}	$\bar{3}m$
Hexagonal	C_6	*6*
	C_{3h}	$\bar{6}$
	C_{6h}	*6/m*
	D_6	*622*
	C_{6v}	*6 mm*
	D_{3h}	$\bar{6}$
	D_{6h}	*6/mmm*

Figure 3.23 shows a crystal plane which passes through lattice points and intersects the axes: $2a$, $3b$, $2c$, where $\vec{a}$, $\vec{b}$, $\vec{c}$ are basic lattice vectors. To obtain Miller indices, we form the ratio $\frac{1}{2} : \frac{1}{3} : \frac{1}{2}$ and put the fractions on the smallest common denominator. The Miller indices are the corresponding numerators. Thus we obtain the Miller indices for the plane: $(hkl) = (323)$.

It also follows that a lattice plane with Miller indices *(hkl)* will be intersected by the axis $\vec{a}$, $\vec{b}$, $\vec{c}$ at distances $\frac{Na}{h}, \frac{Nb}{k}, \frac{Nc}{l}$ where N is an integer. The Miller indices for

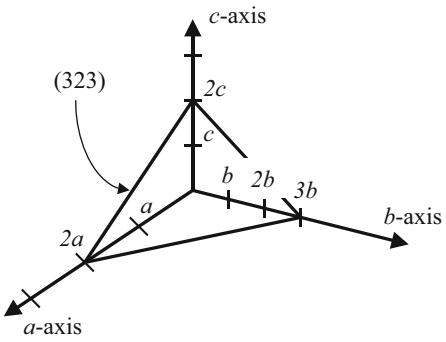


Fig. 3.23 Example of a plane which passes through lattice points. Its Miller indices are $(hkl) = (323)$ and are used to identify this plane in the crystal. These indices are obtained as follows: note where the plane intersects the coordinate axes, it is either an integer multiple or an irreducible fraction of the axis unit length; invert the intercept values; using the appropriate multiplier, convert these inverted values into integer numbers; and enclose the integer numbers in parenthesis

Table 3.3 Conventions used to label directions and planes in crystallography

Notation	Designation
(hkl)	Plane
$\{hkl\}$	Equivalent plane
$[uvw]$	Direction
$<uvw>$	Equivalent direction
$(hkil)$	Plane in hexagonal systems
$[uvtw]$	Direction in hexagonal systems

a few planes in a cubic lattice are shown in Fig. 3.23. These Miller indices are obtained as described above and by using $\frac{1}{1}, \frac{1}{\infty}, \frac{1}{\infty} = 1{:}0{:}0 = (100)$.

For a crystal plane that intersects the origin, one typically has to determine the Miller indices for an equivalent plane which is obtained by translating the initial plane by any lattice vector. The conventions used to label directions and planes in crystallographic systems are summarized in Table 3.3.

The notation for the direction of a straight line passing through the origin is $[uvw]$, where u, v, and w are the three smallest integers whose ratio $u{:}v{:}w$ is equal to the ratio of the lengths (in units of a, b, and c) of the components of a vector directed along the straight line. For example, the symbol for the a-axis in Fig. 3.23, which coincides with vector $\vec{a}$, is [100].

For the indices of both plane and directions, a negative value of the index is written with a bar sign above the index, such as $(\bar{h}kl)$ or $[u\bar{v}w]$.

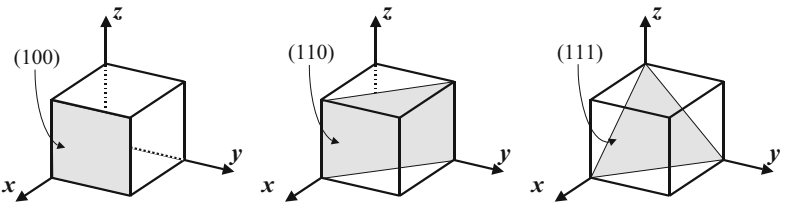
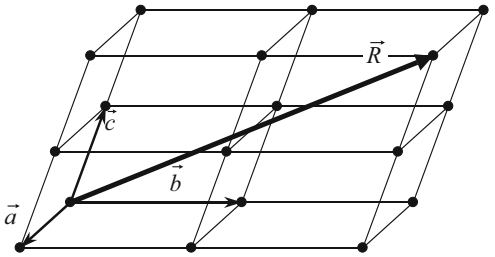


Fig. 3.24 Miller indices of the three principal planes in the cubic structure. If a plane is parallel to an axis, we consider that it "intersects" this axis at infinity and we get the Miller indices: 1,∞,∞ = > 1/1:1/∞:1/∞=1:0:0 = > (100)

Example

Q Determine the direction index for the lattice vector shown below.



A We can decompose the vector $\vec{R}$ as: $\vec{R}= 1\ \vec{a} +2\ \vec{b} +2\ \vec{c}$. This corresponds to $u = 1$, $v = 2$, $w = 2$, and the direction is thus [122].

In cubic systems, such as simple cubic, body-centered cubic, and face-centered cubic lattices, the axes of Fig. 3.24 are chosen to be orthonormal, i.e., the unit vectors are chosen orthogonal and of the same length equal to the side of the cubic unit cell. The axes are then conventionally denoted x, y, and z instead of a, b, and c, as shown in Fig. 3.24.

In addition, for cubic systems, the Miller indices for directions and planes have the following particular and important properties:

- The direction denoted [hkl] is perpendicular to plane denoted (hkl).
- The interplanar spacing is given by the following expression and is shown in the example in Fig. 3.25:

$$d_{hkl} = {}^{a}/_{\sqrt{h^2+k^2+l^2}} \tag{3.4}$$

- The angle θ between two directions [$h_1k_1l_1$] and [$h_2k_2l_2$] is given by the relation:

$$\cos(\theta) = {}^{(h_1h_2+k_1k_2+l_1l_2)}\Big/_{\sqrt{\left(h_1^2+k_1^2+l_1^2\right)\left(h_2^2+k_2^2+l_2^2\right)}} \tag{3.5}$$

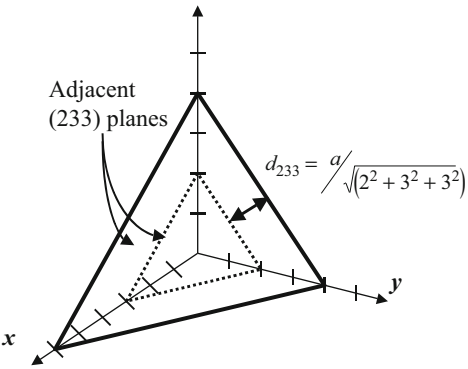


Fig. 3.25 Illustration of the interplanar spacing in a cubic lattice between two adjacent (233) planes

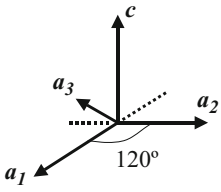
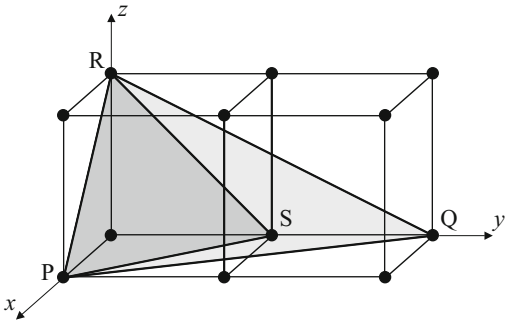


Fig. 3.26 Coordinate axes used to determine Miller indices for hexagonal systems

Example

Q Determine the angle between the two planes shown below (PSR) and (PQR), in a cubic lattice.



A The Miller indices for the (PSR) plane are (111), while they are (212) for the (PQR) plane. The angle θ between these two planes is given by the following cosine function: $\cos(\theta) = \frac{1\times2+1\times1+1\times2}{\sqrt{(1^2+1^2+1^2)(2^2+1^2+2^2)}} = \frac{5\sqrt{3}}{9}$.

The angle between the two planes is therefore 15.8 deg.

In hexagonal systems, the *a*- and *b*-axes of Fig. 3.26 are chosen in the plane formed by the base of the hexagonal unit cell and form a 120 degree angle. They are denoted $\vec{a}_1$ and $\vec{a}_2$ and their length is equal to the side of the hexagonal base. The unit

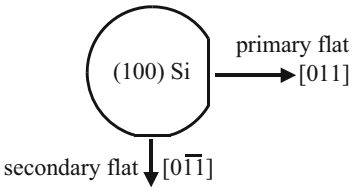


Fig. 3.27 Illustration of the use of primary and secondary flats on a (100) oriented silicon crystal wafer to indicate the in-plane crystallographic orientation of the wafer

vector perpendicular to the base is still denoted c. In addition, it is also conventional to introduce a (redundant) fourth unit vector denoted $\vec{a}_3$ in the base plane and equal to $-(\vec{a}_1 + \vec{a}_2)$, as shown in Fig. 3.26. It is then customary to use a four-index system for planes and directions: (*hkil*) and [*uvtw*], respectively, as shown in Fig. 3.26. The additional index that is introduced for hexagonal systems is such that $i = -(h + k)$ and $t = -(u + v)$, which is a direct consequence of the choice of the fourth unit vector $\vec{a}_3$.

In modern microelectronics, it is often important to know the in-plane crystallographic directions of a wafer and this can be accomplished using Miller indices. During the manufacturing of the circular wafer disk, it is common to introduce a "flat" to indicate a specific crystal direction. To illustrate this, let us consider the (100) oriented silicon wafer shown in Fig. 3.27. A primary flat is such that it is perpendicular to the [011] direction, while a smaller secondary flat is perpendicular to the [01$\bar{1}$] direction.

3.9 Real Crystal Structures

Most semiconductor solids crystallize into a few types of structures which are discussed in this section. They include the diamond, zinc blende, sodium chloride, cesium chloride, hexagonal close-packed, and wurtzite structures.

3.9.1 Diamond Structure

Elements from the column *IV* in the periodic table, such as carbon (the diamond form), germanium, silicon, and gray tin, crystallize in the diamond structure. The Bravais lattice of diamond is face-centered cubic. The basis has two identical atoms located at (0,0,0) and (¼,¼,¼) in the cubic unit cell, for each point of the fcc lattice. The point group of diamond is O_h. The lattice constants are $a = 3.56$, 5.43, 6.65, and 6.46 Å for the four crystals mentioned previously in the same order. The conventional cubic unit cell thus contains eight atoms. There is no way to choose a primitive unit cell such that the basis of diamond contains only one atom.

The atoms which are at least partially in the conventional cubic unit cell are located at the following coordinates: (0,0,0), (0,0,1), (0,1,0), (1,0,0), (1,1,0), (1,0,1), (0,1,1), (1,1,1), (½,½,0), (0,½,½), (½,0,½), (½,½,1), (1,½,½), (½,1,½), (¼,¼,¼), (¾,¾,¼), (¾,¼,¾), and (¼,¾,¾).

The tetrahedral bonding characteristic of the diamond structure is shown in Fig. 3.28a. Each atom has 4 nearest neighbors and 12 s nearest neighbors. For example, the atom located at (¼,¼,¼) at the center of the cube in Fig. 3.28b has four

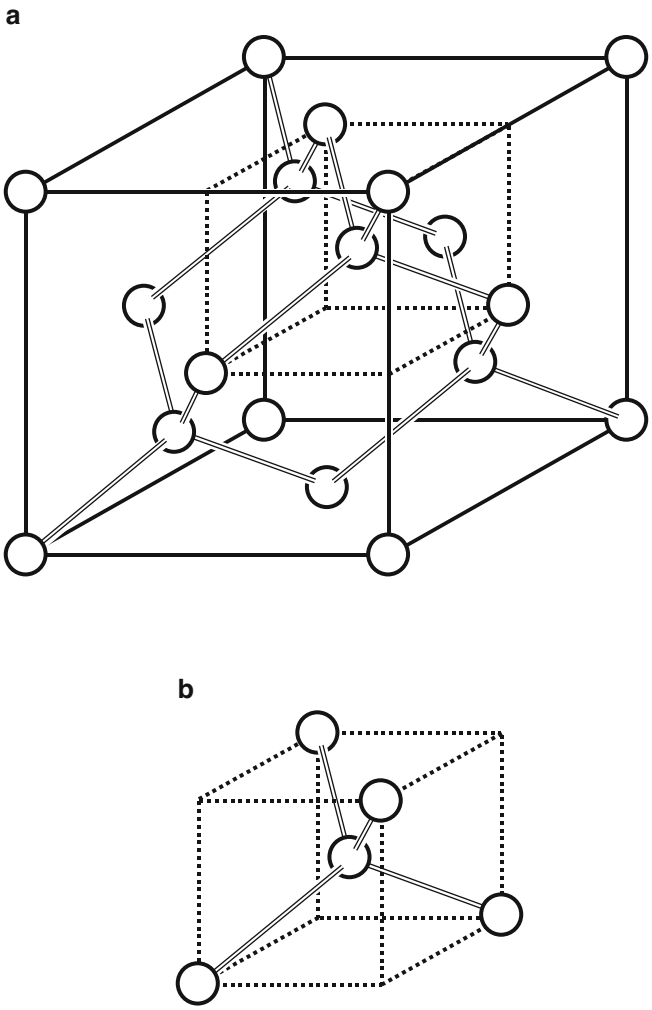


Fig. 3.28 (**a**) Diamond lattice. The Bravais lattice is face-centered cubic with a basis consisting of two identical atoms displaced from each other by a quarter of the cubic body diagonal. The atoms are connected by covalent bonds. The cube outlined by the dashed lines shows one tetrahedral unit. (**b**) Tetrahedral unit of the diamond lattice

nearest neighbors also shown in Fig. 3.28b which are located at (0,0,0), (½,½,0), (0,½,½), and (½,0,½).

The number of atoms/unit cell for the diamond lattice is found from $n_i = 4$, $n_f = 6$, and $n_c = 8$ where n_i, n_f, and n_c are the numbers of points in the interior, on faces, and on corners of the cubic unit cell shown in Fig. 3.28a, respectively. Note that each of the n_f points is shared between two cells and each of the n_c points is shared between eight cells. Therefore: $n_u = 4 + \frac{6}{6} + \frac{8}{8} = 8$ atoms/unit cell. The atomic density or the number of atoms per cm^3, n, is given by $n = \frac{n_u}{a^3}$ atoms/unit cell. For example, for silicon, we have $a = 5.43$ Å, and $n = 8/(0.543 \times 10^{-7})^3 = 5 \times 10^{22}$ atoms/cm^3.

3.9.2 Zinc Blende Structure

The most common crystal structure for III–V compound semiconductors, including GaAs, GaSb, InAs, and InSb, is the sphalerite or zinc blende structure shown in Fig. 3.29. The point group of the zinc blende structure is T_d.

The zinc blende structure has two different atoms. Each type of atom forms a face-centered cubic lattice. Each atom is bounded to four atoms of the other type. The sphalerite structure as a whole is treated as a face-centered cubic Bravais lattice with a basis of two atoms displaced from each other by $(a/4)(x + y + z)$, i.e., one fourth of the length of a body diagonal of the cubic lattice unit cell. Some important properties of this crystal result from the fact that the structure does not appear the same when viewed along a body diagonal from one direction and then the other. Because of this, the sphalerite structure is said to lack inversion symmetry. The crystal is therefore polar in its <111> directions, i.e., the [111] and the $[\bar{1}\,\bar{1}\,\bar{1}]$ directions are not equivalent. When both atoms are the same, the sphalerite structure has the diamond structure, which has an inversion symmetry and was discussed previously.

In the case of GaAs, for example, the solid spheres in Fig. 3.29 represent Ga atoms and the open spheres represent As atoms. Their positions are:

Ga: (0,0,0), (½,½,0), (0,½,½), (½,0,½), (½,1,½), (½,½,1), (1,½,½)
As: (¼,¼,¼), (¾,¾,¼), (¾,¼,¾), (¼,¾,¾)

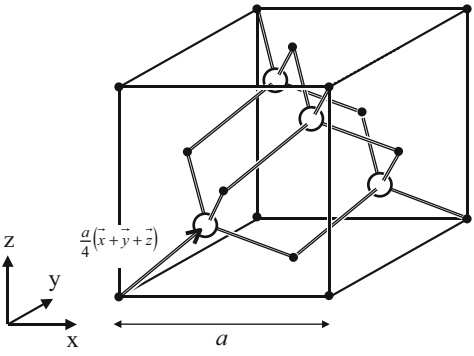


Fig. 3.29 Cubic unit cell for the zinc blende structure. The Bravais lattice is face-centered cubic with a basis of two different atoms represented by the open and solid spheres and separated by a quarter of the cubic body diagonal. The crystal does not appear the same when viewed along a body diagonal from one direction or the other

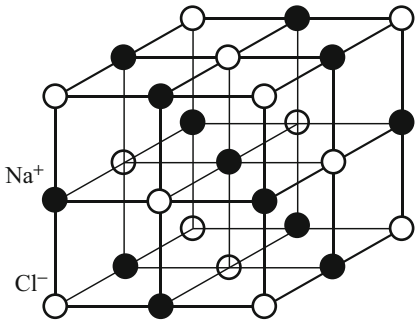


Fig. 3.30 Sodium chloride crystal. The Bravais lattice is face-centered cubic with a basis of two ions: one Cl^- ion at (0,0,0) and one Na^+ ion at (½,½,½), separated by one half of the cubic body diagonal. The figure shows one cubic unit cell

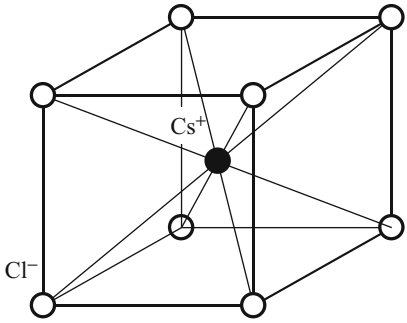


Fig. 3.31 The cesium chloride crystal structure. The Bravais lattice is cubic with a basis of two ions: one Cl^- ion at (0,0,0) and one Cs^+ ion at (½,½,½), separated by one half the cubic body diagonal

3.9.3 Sodium Chloride Structure

The structure of sodium chloride, NaCl, is shown in Fig. 3.30. The Bravais lattice is face-centered cubic and the basis consists of one Na atom and one Cl atom separated by one half the body diagonal of the cubic unit cell. The point group of the sodium chloride structure is O_h.

There are four units of NaCl in each cubic unit cell, with atoms in the positions:

Cl: (0,0,0), (½,½,0), (½,0, ½), (0, ½,½)
Na: (½,½,½), (0,0, ½), (0, ½,0), (½,0,0)

3.9.4 Cesium Chloride Structure

The cesium chloride structure is shown in Fig. 3.31. The Bravais lattice is simple cubic and the basis consists of two atoms located at the corner (0,0,0) and center positions (½,½,½) of the cubic unit cell. Each atom may be viewed as at the center of a cube of atoms of the opposite kind, so that the number of nearest neighbors or coordination number is eight. The point group of the cesium chloride structure is T_d.

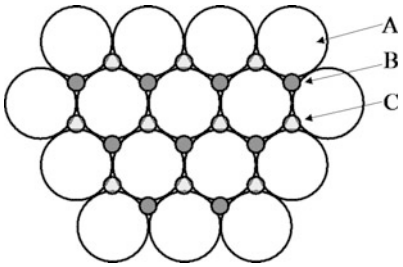


Fig. 3.32 The closed-packed array of spheres. Note the three different possible positions, A, B, and C for the successive layers. The most space-efficient way to arrange identical spheres or atoms in a plane is to first place each sphere in contact with six others in that plane (positions A). The most stable way to stack a second layer of such spheres is by placing each one of them in contact with three spheres of the bottom layer (positions B). The third stable layer can then either be such that the spheres occupy positions above A or C

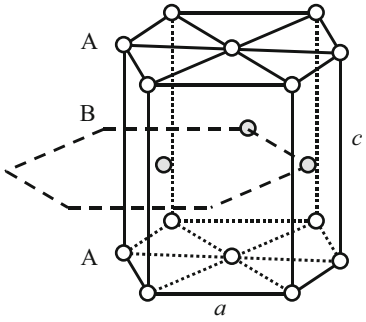


Fig. 3.33 The hexagonal close-packed (hcp) structure. This Bravais lattice of this structure is hexagonal, with a basis of two identical atoms. It is constructed by stacking layers in the ABABAB... sequence. The lattice parameters a and c are indicated

3.9.5 Hexagonal Close-Packed Structure

The simplest way to stack layers of spheres is to place centers of spheres (atoms) directly above one another. The resulting structure is called a *simple hexagonal structure*. There is, in fact, no example of crystals with this structure because it is unstable. However, spheres can be arranged in a single hexagonal close-packed layer A (Fig. 3.32) by placing each sphere in contact with six others. A second similar layer B may be added by placing each sphere of B in contact with three spheres of the bottom layer, at positions B in Fig. 3.32. This arrangement has the lowest energy and is therefore stable. A third layer may be added in two different ways. We obtain the cubic structure if the spheres of the third layer C are added over the holes in the first layer A that are not occupied by B, as in Fig. 3.32. We obtain the hexagonal close-packed structure (Fig. 3.33) when the spheres in the third layer are placed directly over the centers of the spheres in the first layer, thus replicating layer A. The Bravais lattice is hexagonal. The point group of the hexagonal close-packed structure is D_{6h}. The fraction of the total volume occupied by the spheres is 0.74 for both structures (see Problems).

Table 3.4 c/a parameter for various hexagonal crystals

Crystal	*c/a*
Be	1.581
Mg	1.623
Ti	1.586
Zn	1.861
Cd	1.886
Co	1.622
Y	1.570
Zr	1.594
Gd	1.592

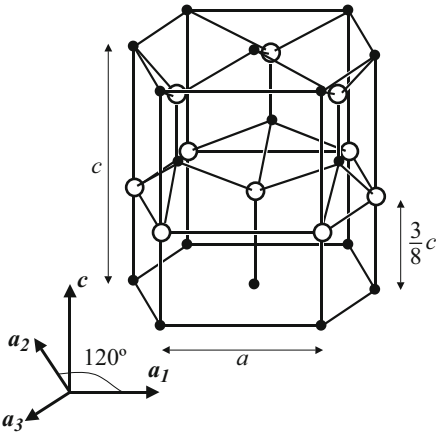


Fig. 3.34 The wurtzite structure consists of two interpenetrating hcp structures, each with a different atom, shifted along the c-direction. The bonds between atoms and the hexagonal symmetry are shown

Zinc, magnesium, and low-temperature form of titanium have the hcp structure. The ratio *c/a* for ideal hexagonal close-packed structure in Fig. 3.33 is 3.633. The number of nearest-neighbor atoms is 12 for hcp structures. Table 3.4 shows the *c/a* parameter for different hexagonal crystals.

3.9.6 Wurtzite Structure

A few III–V and several II–VI semiconductor compounds have the wurtzite structure shown in Fig. 3.34.

This structure consists of two interpenetrating hexagonal close-packed lattices, each with different atoms, ideally displaced from each other by 3/8*c* along the *z*-axis. There is no inversion symmetry in this crystal, and polarity effects are observed along the *z*-axis. The Bravais lattice is hexagonal with a basis of four atoms, two of each kind. The point group of the wurtzite structure is C_{6v}.

3.9.7 Packing Factor

The packing factor is the maximum proportion of the available volume in a unit cell that can be filled with hard spheres. Let us illustrate this concept with a few examples.

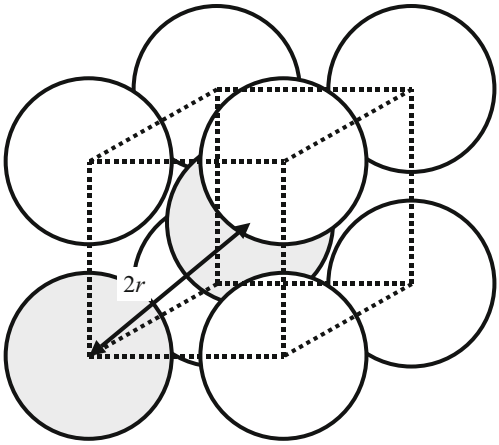
For a simple cubic lattice, the center-to-center distance between the nearest atoms is a. So the maximum radius of the atom is $a/2$. Since there is only one atom point per cubic unit cell in this case, the packing factor is $\frac{\frac{4}{3}\pi\left(\frac{a}{2}\right)^3}{a^3} = 0.52$.

The following two examples illustrate the determination of the packing factor for the other two cubic lattices.

Example

Q Determine the packing factor for a body-centered cubic lattice.

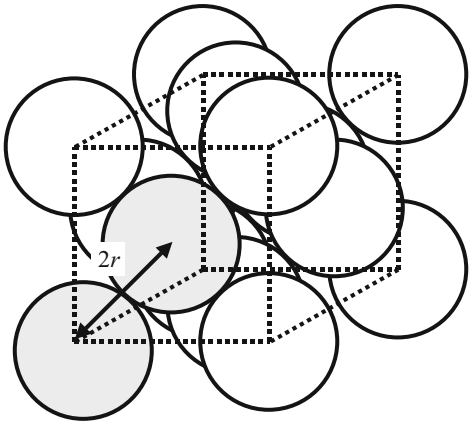
A Let us consider the bcc lattice shown in the figure below and an atom located at one corner of the cubic unit cell. Its nearest neighbor is an atom which is located at the center of the cubic unit cell and which is at a distance of $\frac{\sqrt{3}}{2}a$ where a is the side of the cube. The maximum radius r for the atoms is such that these two atoms touch and therefore $2r = \frac{\sqrt{3}}{2}a$. There are two atoms in a bcc cubic unit cell, so the maximum volume filled by the spheres is $2 \times \frac{4\pi}{3}\left(\frac{\sqrt{3}}{4}a\right)^3$. The packing factor is calculated by taking the ratio of the total sphere volume to that of the unit cell and yields $\frac{2\times\frac{4\pi}{3}\left(\frac{\sqrt{3}}{4}a\right)^3}{a^3} = \frac{\pi\sqrt{3}}{8} = 0.68$.



Example

Q Determine the packing factor for a face-centered cubic lattice.

A Let us consider the fcc lattice shown in the figure below and an atom located at one corner of the cubic unit cell. Its nearest neighbor is an atom which is located at the center of an adjacent face of the cubic unit cell and which is at a distance of $\frac{\sqrt{2}}{2}a$ where a is the side of the cube. The maximum radius r for the atoms is such that these two atoms touch and therefore $2r = \frac{\sqrt{2}}{2}a$. There are four atoms in a fcc cubic unit cell, so the maximum volume filled by the spheres is $4 \times \frac{4\pi}{3}\left(\frac{\sqrt{2}}{4}a\right)^3$. The packing factor is calculated by taking the ratio of the total sphere volume to that of the unit cell and yields: $\frac{4\times\frac{4}{3}\pi\left(\frac{\sqrt{2}}{4}a\right)^3}{a^3} = 0.7405$.



The diamond structure has the face-centered cubic structure with a basis of two identical atoms. The packing factor of diamond structure is only 46 percent of that in the fcc structure, so diamond structure is relatively empty (see Problems).

3.10 The Reciprocal Lattice

When we have a periodic system, one lattice point is equivalent to another lattice point, so we expect a simple relation to exist between physical quantities at these respective lattice points. Consider, for example, the local density of charge $\rho(\vec{r})$. We should expect this quantity to have the same periodicity as the lattice. But it is mathematically known that any periodic function can be expanded into a Fourier series. In a crystal lattice, all physical quantities have the periodicity of the lattice, in all directions. Let us consider the above physical quantity $\rho(\vec{r})$. From now, we will

use a three-dimensional formalism. This function is periodic and can be expanded into a Fourier series:

$$\rho(\vec{r}) = \sum_{\vec{K}} \mathrm{P}(\vec{K}) \exp(i\,\vec{K}\vec{r}) \tag{3.6}$$

where the vector $\vec{K}$ is used to index the summation and the Fourier coefficients $\mathrm{P}(\vec{K})$. This vector $\vec{K}$ has the dimension of an inverse distance and, for a periodic function, can take on discrete values in a three-dimensional sum. Let us now express that the function $\rho(\vec{r})$ is periodic by calculating its value after displacement by a lattice vector $\vec{R}$:

$$\rho(\vec{r}) = \rho(\vec{r} + \vec{R}) = \sum_{\vec{K}} \mathrm{P}(\vec{K}) \exp\left[i\,\vec{K} \cdot (\vec{r} + \vec{R})\right] \tag{3.7}$$

which becomes

$$\sum_{\vec{K}} \mathrm{P}(\vec{K}) \exp(i\,\vec{K} \cdot \vec{r}) = \sum_{\vec{K}} \mathrm{P}(\vec{K}) \exp\left[i\,\vec{K} \cdot (\vec{r} + \vec{R})\right] \tag{3.8}$$

has to be satisfied for any given function which is periodic with the periodicity of the lattice. This can be satisfied if and only if

$$\exp\left[i\,\vec{K} \cdot (\vec{r} + \vec{R})\right] = \exp(i\,\vec{K} \cdot \vec{r})$$

or

$$\exp(i\,\vec{K} \cdot \vec{R}) = 1 \tag{3.9}$$

for any lattice vector Eq. (3.9) is the major relation which allows us to introduce the so-called reciprocal lattice which is spanned by the vectors $\vec{K}$. What follows next is a pure mathematical consequence of Eq. (3.9) which is equivalent to

$$\vec{K} \cdot \vec{R} = 2\pi m \tag{3.10}$$

where $m = 0, \pm 1, \pm 2, \ldots$ is an integer. Using the expression for $\vec{R}$ from Eq. (3.1) of Chap. 3, we obtain

$$(\vec{K} \cdot \vec{a})n_1 + (\vec{K} \cdot \vec{b})n_2 + (\vec{K} \cdot \vec{c})n_3 = 2\pi m \tag{3.11}$$

where n_1, n_2, and n_3 are arbitrary integers which come from the choice of the vector $\vec{R}$. Because the sum of three terms is an integer if and only if each term itself is integer leads us to

$$\begin{cases} \vec{K} \cdot \vec{a} = 2\pi h_1 \\ \vec{K} \cdot \vec{b} = 2\pi h_2 \\ \vec{K} \cdot \vec{c} = 2\pi h_3 \end{cases} \text{ with } h_{1,2,3} = 0; \pm 1; \pm 2, \ldots \tag{3.12}$$

Here, $h_{1,\,2,\,3}$ is not related to Planck's constant.

Let us now define three basis vectors ($\vec{A}$, $\vec{B}$, $\vec{C}$) in order to express $\vec{K}$ in the same way as we did it for real lattice vectors in Eq. (3.12) of Chap. 3. These basis vectors define what we call the reciprocal lattice. Any reciprocal lattice vector $\vec{K}$ can thus be represented as

$$\vec{K} = h_1 \vec{A} + h_2 \vec{B} + h_3 \vec{C}; \tag{3.13}$$

From (3.12) and (3.11), we have

$$\begin{cases} (\vec{A} \cdot \vec{a}) h_1 + (\vec{B} \cdot \vec{a}) h_2 + (\vec{C} \cdot \vec{a}) h_3 = 2\pi h_1 \\ (\vec{A} \cdot \vec{b}) h_1 + (\vec{B} \cdot \vec{b}) h_2 + (\vec{C} \cdot \vec{b}) h_3 = 2\pi h_2 \\ (\vec{A} \cdot \vec{c}) h_1 + (\vec{B} \cdot \vec{c}) h_2 + (\vec{C} \cdot \vec{c}) h_3 = 2\pi h_3 \end{cases} \tag{3.14}$$

Equation (3.14) can be satisfied only when

$$\begin{cases} \vec{A} \cdot \vec{a} = \vec{B} \cdot \vec{b} = \vec{C} \cdot \vec{c} = 2\pi \\ \text{and} \\ \vec{A} \cdot \vec{b} = \vec{A} \cdot \vec{c} = 0 \\ \vec{B} \cdot \vec{a} = \vec{B} \cdot \vec{c} = 0 \\ \vec{C} \cdot \vec{b} = \vec{C} \cdot \vec{a} = 0 \end{cases} \tag{3.15}$$

Equation (3.15) defines the relation between the direct ($\vec{a}$, $\vec{b}$, $\vec{c}$) and reciprocal ($\vec{A}$, $\vec{B}$, $\vec{C}$) basis lattice vectors and gives the means to construct ($\vec{A}$, $\vec{B}$, $\vec{C}$) from ($\vec{a}$, $\vec{b}$, $\vec{c}$):

$$\begin{cases} \vec{A} = 2\pi \dfrac{\vec{b} \times \vec{c}}{\vec{a} \cdot \left(\vec{b} \times \vec{c}\right)} \\ \vec{B} = 2\pi \dfrac{\vec{c} \times \vec{a}}{\vec{a} \cdot \left(\vec{b} \times \vec{c}\right)} \\ \vec{C} = 2\pi \dfrac{\vec{a} \times \vec{b}}{\vec{a} \cdot \left(\vec{b} \times \vec{c}\right)} \end{cases} \tag{3.16}$$

These relations are a natural consequence of vector algebra in three dimensions. The volumes that these basis vectors define in the real and reciprocal lattices satisfy the relation (see Problems):

$$\vec{A} \cdot \left(\vec{B} \times \vec{C}\right) = \frac{8\pi^3}{\vec{a} \cdot \left(\vec{b} \times \vec{c}\right)} \tag{3.17}$$

We note that the vectors of reciprocal space have the same dimensions as the wavenumbers and momenta of electromagnetic waves. We also note the direct lattice is the reciprocal of its own reciprocal lattice. The concept of reciprocal or momentum space turns out to be extremely important for the classification of electron states in a crystal in quantum theory.

3.11 The Brillouin Zone

In the reciprocal lattice, we can construct unit cells as we did for the real lattice earlier in this chapter. The construction of the Wigner-Seitz cell in the reciprocal lattice follows the same rules as in the real lattice and gives the smallest unit cell in *k*-space called the "first Brillouin zone" and shown in Fig. 3.10. Draw the perpendicular bisector planes of the translation vectors from the chosen center to the nearest equivalent sites in the reciprocal lattice, and you have formed the first Brillouin zone.

3.12 Summary

In this chapter, the structure of crystals has been described. The concepts of Bravais lattice, crystal systems, unit cell, point groups, space groups, Miller indices, and packing factor have been introduced. The symmetry properties of crystals have been discussed. The most common crystal structures for semiconductors have been described. We have also introduced the concept of the reciprocal lattice. We have shown that for every periodic lattice in real space $\vec{R}$, it is possible to construct a

periodic reciprocal lattice in $\vec{K}$ space. The reciprocal lattice is the lattice in so-called momentum space. The Wigner Seitz cell of the reciprocal lattice is called the first Brillouin zone.

Problems

1. Figure 3.6 illustrates the definition of the angles and unit cell dimensions of the crystalline material. If a unit cell has a characteristic of $a = b = c$ and $\alpha = \beta = \gamma = 90^\circ$, it forms a cubic crystal system, which is the case of Si and GaAs.
   (a) How many Bravais lattices are classified in the cubic system?
   (b) Draw simple three-dimensional unit cells for each Bravais lattice in the cubic system.
   (c) How many lattice points are contained in the unit cell for each Bravais lattice in the cubic system?
2. Draw the four Bravais lattices in orthorhombic lattice system.
3. Show that the C_5 group is not a crystal point group. In other words, show that, in crystallography, a rotation about an axis and through an angle $\theta = \frac{2\pi}{5}$ cannot be a crystal symmetry operation.
4. Determine if the plane (111) is parallel to the following directions: [100], $[\bar{2}11]$, and $[\bar{1}\bar{1}0]$.
5. For cesium chloride, take the fundamental lattice vectors to be $\vec{a} = a\,\vec{x}$, $\vec{b} = a\,\vec{y}$, and $\vec{c} = a(\vec{x} + \vec{y} + \vec{z})$. Describe the parallelepiped unit cell and find the cell volume.
6. GaAs is a typical semiconductor compound that has the zinc blende structure.
   (a) Draw a cubic unit cell for the zinc blende structure showing the positions of Ga and As atoms.
   (b) Make a drawing showing the in-plane crystallographic directions and the positions of the atoms for the (111) lattice plane.
   (c) Repeat for the (100) plane.
   (d) Calculate the surface density of atoms in (100) plane.
7. (a) What are the interplanar spacings d for the (100), (110), and (111) planes of Al ($a = 4.05$ Å)?
   (b) What are the Miller indices of a plane that intercepts the x-axis at a, the y-axis at $2a$, and the z-axis at $2a$?
8. Show that the c/a ratio for an ideal hexagonal close-packed structure is $(8/3)^{1/2} = 1.633$. If c/a is significantly larger than this value, the crystal structure may be thought of as composed of planes of closely packed atoms, the planes being loosely stacked.
9. Show that the packing factor in a hexagonal close-packed structure is 0.74.
10. Show that the packing factor for the diamond structure is 46% of that in the fcc structure.

11. Let ($\vec{a}$, $\vec{b}$, $\vec{c}$) be a basis lattice vectors for a direct lattice and ($\vec{A}$, $\vec{B}$, $\vec{C}$) be the basis lattice vectors for the reciprocal lattice defined by Eq. (3.16). Prove that the volume defined by these vectors is given by $\vec{A} \cdot \left(\vec{B} \times \vec{C}\right) = \dfrac{8\pi^3}{\vec{a} \cdot \left(\vec{b} \times \vec{c}\right)}$.

Further Reading

Dalven R (1990) Introduction to applied solid state physics: topics in the applications of semiconductors, superconductors, ferromagnetism, and the nonlinear optical properties of solids. Plenum Press, New York
Holden A (1992) The nature of solids. Dover, New York
Kittel C (1986) Introduction to solid state physics. Wiley, New York
Loretto MH (1994) Electron-beam analysis of materials. Chapman & Hall, London
Lovett DR (1999) Tensor properties of crystals. Institute of Physics, Bristol
Mayer JW, Lau S (1990) Electronic materials science for integrated circuits in Si and GaAs. Macmillan, New York
McKelvey JP (1966) Solid state and semiconductor physics. Harper and Row, New York
Pierret RF (1989) Advanced semiconductor fundamentals. Addison-Wesley, Reading
Rhodes G (1993) Crystallography made crystal clear. Academic Press, San Diego
Rosenberg HM (1978) The solid state, Oxford physics series. Clarendon Press, Oxford
Scott WR (1964) Group theory. Dover Publications, New York
Weyl H (1950) The theory of groups and quantum mechanics. Dover Publications, New York
Wolfe CM, Holonyak N, Stillman GE (1989) Physical properties of semiconductors. Prentice-Hall, Englewood Cliffs
Yu PY, Cardona M (1999) Fundamentals of semiconductors: physics and materials properties. Springer, New York
Ziman JM (1969) Elements of advanced quantum theory. Cambridge University Press, London
Ziman JM (1998) Principles of the theory of solids. Cambridge University Press, Cambridge

Introduction to Quantum Mechanics

4

4.1 The Quantum Concepts

In Chapter 1 we saw that classical mechanics was incapable of explaining the optical spectra emitted by atoms or even the existence of atoms. Bohr developed a model for the atom of hydrogen by assuming the quantization of the angular momentum, which was an introduction to wave or quantum mechanics. Quantum mechanics is a more precise approach to describe nearly all physical phenomena which reduces to classical mechanics in the limit where the masses and energies of the particles are large or macroscopic.

In this section, we will illustrate the success of quantum mechanics through the historically important examples of blackbody radiation, wave-particle duality, the photoelectric effect, and the Davisson and Germer experiment.

4.1.1 Blackbody Radiation

As introduced in Chap. 1, a blackbody is an ideal source of electromagnetic radiation, and the radiated power dependence was depicted as a function of wavelength in Fig. 1.3 for several temperatures of the blackbody.

When the temperature of the body is at or below room temperature, the radiation is mostly in the infrared spectral region, i.e., not detectable by the human eye. When the temperature is raised, the emission power increases, and its peak shifts toward shorter wavelengths as shown in Fig. 1.3. Several attempts to explain this observed blackbody spectrum were made using classical mechanics in the latter half of the nineteenth century, and one of the most successful ones was proposed by Rayleigh and Jeans.

In their classical model, a solid at thermal equilibrium is seen as consisting of vibrating atoms which are considered harmonic electric oscillators which generate standing waves, or modes, through reflections within the cavity. A *continuous*

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_4

spectrum of vibrational mode frequencies $\nu = \omega/2\pi = c/\lambda$ where c denotes the velocity of light and λ the wavelength of the oscillations. These atomic vibrations cause the emission of electromagnetic radiation in a continuous frequency range too. To determine the power radiated, one has to first determine the energy distribution for each frequency. According to the classical law of equipartition of energy, the average energy per degree of freedom for a blackbody in equilibrium is equal to k_bT, where k_b is the Boltzmann constant ($k_b = 8.614 \times 10^{-5}$ eV·K^{-1}) and T the absolute temperature in degrees K. The number of modes per unit volume is the number of degrees of freedom for an electromagnetic radiation.

To calculate this number, a simple model can be used which involves propagating waves in a rectangular box. Only certain frequencies of waves are allowed as a result of boundary conditions at the limits of the box. In addition, there are two possible polarization directions for the waves, corresponding to what are called "TE" and "TM" propagation modes. The total number of modes per unit volume and per unit frequency interval is $\frac{8\pi\nu^2}{c^3}$. Therefore, the distribution of energy radiated by a blackbody per unit volume and per unit frequency interval is $u(\nu, T) = \frac{8\pi\nu^2}{c^3}k_bT$. Considering that this energy is radiated at the speed of light, and by expressing this distribution in terms of wavelength, we get the distribution of power radiated per unit area and per unit wavelength interval as $w(\lambda, T) = \frac{8\pi c}{\lambda^4}k_bT$. Both expressions $u(\nu, T)$ and $w(\lambda, T)$ are called the Rayleigh-Jeans law. This law is illustrated by a dashed line in Fig. 1.3 for $T = 2000$ K. It shows that this classical theory was in reasonably good agreement with experimental observations at longer wavelengths. However, over the short-wavelength portion of the spectrum, there was significant divergence between experiment and theory. This is because we assumed the classical law of equipartition of energy was valid at all wavelengths. This discrepancy came to be known as the "ultraviolet catastrophe" because the integration of the Rayleigh-Jeans law over all frequencies or wavelengths would theoretically lead to an infinite amount of radiated power.

These experimental observations could therefore not be explained until 1901, when Max Planck provided a detailed theoretical explanation of the observed blackbody spectrum by introducing the hypothesis that the atoms vibrating at a frequency v in a material could only radiate or absorb energy in discrete or *quantized* packages proportional to the frequency:

$$E_n = nh\nu = n\hbar\omega \qquad n = 0, 1, 2, \ldots \tag{4.1}$$

where n is an integer used to express the quantization, h is Planck's constant, and $\hbar = h/2\pi$ is the reduced Planck's constant, obtained by matching theory to experiment and is called Planck's constant. This also means that the energy associated with each mode of the radiated electromagnetic field at a frequency ν did not vary continuously (with an average value kT) but was an integral multiple of $h\nu$. Planck then made use of the Boltzmann probability distribution to calculate the average energy associated with each frequency mode. This Boltzmann distribution states that the probability for a system in equilibrium at temperature T to have an energy E is proportional to $e^{-E/kT}$ and can be expressed as:

$$P(E_n) = \frac{e^{-E_n/k_bT}}{\sum\limits_E e^{-E/k_bT}} \tag{4.2a}$$

and is normalized because the total probability after summation over all possible values of E has to be unity. Taking into account the quantization condition in Eq. (4.1), the average energy <E> associated with each frequency mode ν can thus be written as:

$$\langle E \rangle = \sum_{E_n} E_n P(E_n) = \frac{\sum\limits_{n=0}^{\infty} (nh\nu) e^{-nh\nu/k_bT}}{\sum\limits_{n=0}^{\infty} e^{-nh\nu/k_bT}} = \frac{h\nu}{e^{h\nu/kT} - 1}. \tag{4.2b}$$

Therefore, after multiplying by the number of modes per unit volume and frequency $\frac{8\pi\nu^2}{c^3}$, we obtain the distribution of energy radiated by a blackbody at frequency of ν in this model:

$$u(\nu, T) = \frac{8\pi\nu^2}{c^3} \frac{h\nu}{e^{h\nu/k_bT} - 1} \tag{4.2c}$$

This expression is found to be in good agreement with experimental observations. Actually, there is apparently no other physical law which fits experiments with a higher degree of precision. In the limit of small frequencies, or long wavelengths, this relation simplifies into the Rayleigh-Jeans law because we can make the approximation:

$$e^{h\nu/kT} - 1 \approx {}^{h\nu}\!/_{k_bT}$$

We can thus see that the classical equipartition law is no longer valid whenever the frequency is not small compared with k_bT/h. Moreover, this expression shows that high-frequency modes have very small average energy.

This example of the blackbody radiation already shows that, for atomic dimension systems, the classical view which always allows a continuum of energies is incorrect. Discrete steps in energy, or energy quantization, must occur and is a central feature of the quantum approach to real-life phenomena.

4.1.2 The Photoelectric Effect

In 1902, Philipp Lenard studied the emission of electrons from a metal under illumination. And, in particular, he studied how their energy varied with the intensity and the frequency of the light.

A simplified setup of his experiment is schematically depicted in Fig. 4.1. It involved a chamber under vacuum, two parallel metal plates on which a voltage was applied. Light was shone onto a metal plate. The electrons in it were then excited by

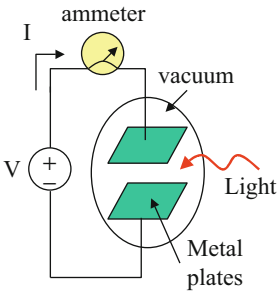


Fig. 4.1 Simplified experimental setup used by Lenard. A chamber in vacuum contains two parallel metal plates on which a voltage is applied. Light shining onto a metal plate gives enough energy to the electrons of the plate to make them leave the plate and be accelerated by the electric field

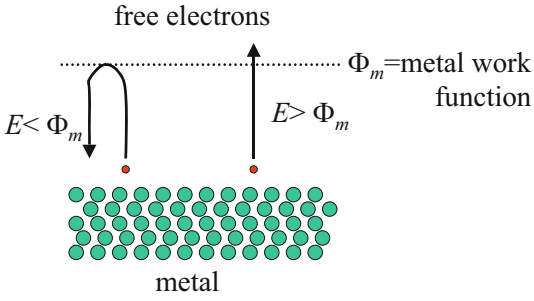


Fig. 4.2 The work function of a metal, denoted Φ_m, is the minimum amount of energy that an electron needs to acquire to leave the metal

this incident light and could gain enough energy to leave the metal surface into the vacuum. This was called the photoelectric effect. These electrons can then be accelerated by the electric field between the metal plate and reach the opposite plate, thus leading to an electrical current that can be measured using a sensitive ammeter.

It was known at the time that there existed a minimum energy, called the metal work function and denoted by Φ_m, which was required to have an electron break free from a given metal, as illustrated in Fig. 4.2. One had to give an energy $E > \Phi_m$ to an electron in order to enable it to escape the attraction of the metal ions.

Example

Q: In the photoelectric effect, the stopping potential V_0, which is the potential required to bring the emitted photoelectrons to rest, can be experimentally determined. This potential is related to the work function Φ_m through $qV_0 = \frac{hc}{\lambda} - \Phi_m$, where λ is the wavelength of the incident photon. For a photon with a wavelength of 2263 Å, incident on the surface of lithium, we experimentally find $V_0 = 3.00$ V. Determine the work function of Li.

A: Using the above formula, we get:

$$\begin{aligned}\Phi_m &= \frac{hc}{\lambda} - qV_0 \\ &= \frac{\left(6.62617 \times 10^{-34}\right)\left(2.99792 \times 10^{8}\right)}{2263 \times 10^{-10}} - \left(1.60218 \times 10^{-19}\right)(3) \\ &= 3.97 \times 10^{-19}\,\text{J} \\ &= 2.48\,\text{eV}\end{aligned}$$

As his light source, Lenard used a carbon-arc lamp emitting a broad range of frequencies and was able to increase its total intensity a thousandfold. With such a powerful arc lamp, it was then possible to obtain monochromatic light at various arbitrary frequencies and each with reasonable power. Lenard could then investigate the photoelectric effect when the frequency of the incident light was varied. To his surprise, he found that below a certain frequency (i.e., certain color), no current could be measured, suggesting that the electrons could not leave the metal any more even when he increased the intensity of light by several orders of magnitude.

In 1905, Albert Einstein successfully interpreted Lenard's results by simply assuming that the incident light was composed of indivisible quanta or packets of energy, each with an energy equal to $h\nu$ where h is Planck's constant and ν is a frequency. He called each quantum a photon. The electrons in the metal could then receive an energy E equal to that of a quantum of light or a photon, i.e., $E = h\nu$. Therefore, if the frequency ν was too low, such that $E = h\nu$ was smaller than Φ_m, the electrons would not have enough energy to escape the metal plate, independently of how high the intensity of light was, as shown in Fig. 4.3. However, if the frequency was high enough, such that $E = h\nu$ was higher than Φ_m, electrons could escape the metal. Albert Einstein won the Nobel Prize in Physics in 1921 for his work on the photoelectric effect.

It is interesting to know that an American experimental physicist, Robert Millikan, who did not accept Einstein's theory, worked for 10 years to show its failure. In spite of all his efforts, he found a rather disappointing result as he ironically confirmed Einstein's theory by measuring Planck's constant to within 0.5%. One consolation was that he did get awarded the Nobel Prize in Physics in 1923 for his experiments!

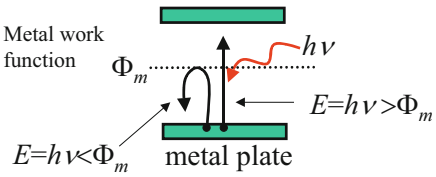


Fig. 4.3 Schematic diagram of the escape mechanism of an electron in the metal plate receiving a photon with energy $h\nu$. If the photon energy is lower than the work function, the electron does not escape. If the photon energy is higher than the work function, the electron receives enough energy to reach the vacuum level and leave the metal

4.1.3 Wave-Particle Duality

The previous discussions on the Bohr atom in Chap. 1, the blackbody radiation and the photoelectric effect, led to the conclusion that the electromagnetic radiation has a quantum nature because it exhibits particle-like properties.

In 1925, Louis de Broglie conjectured that, since the electromagnetic radiation had particle-like properties, particles (e.g., electrons) should have wave-like properties as well. This was called the wave-particle duality. He postulated that a particle with a momentum p can be viewed as a wave with a wavelength given by:

$$\lambda = \frac{h}{p} \tag{4.3}$$

This relation establishes the relationship between a particle and a wave in nature. This concept, as well as the others introduced in the previous examples, clearly proves that classical mechanics was limited and that a new theory was required which would take into account the quantum structure of matter, electromagnetic fields, and the wave-particle duality. In 1927 such a theory was created and called wave or quantum mechanics (Liboff 1998; Davydov 1965).

4.1.4 The Davisson-Germer Experiment

The first complete and convincing evidence of de Broglie's hypothesis came from an experiment that Clinton Davisson and Lester Germer did at the Bell Laboratories in 1926. Using an electron gun, they directed beams of electrons onto a nickel crystal plate from where they were then reflected, as schematically depicted in Fig. 4.4. A sensitive screen, such as a photographic film, was put above the nickel target to get information on the directions in which the electrons reflected most. On it, they observed concentric circular rings, showing that the electrons were more likely to appear at certain angles than others. This was similar to a diffraction pattern and confirmed that these electrons had a wave-like behavior.

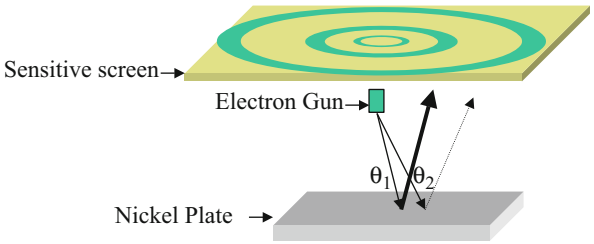


Fig. 4.4 Schematic of the experimental setup in the Davisson-Germer experiment. A beam of electrons is directed on a nickel plate from which the electrons are reflected. They then hit a sensitive screen and create a ring pattern

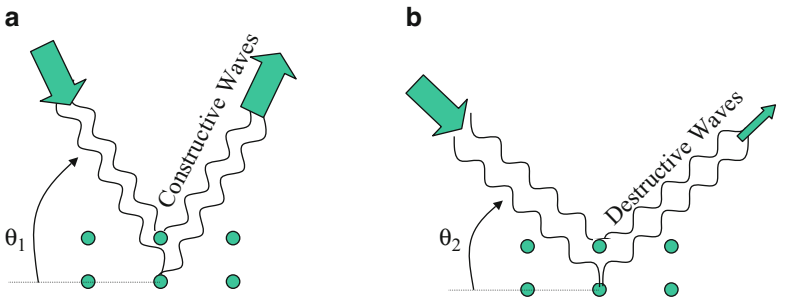


Fig. 4.5 (**a**) Constructive diffraction and (**b**) destructive diffraction condition for the waves reflected from a crystal surface. In the constructive diffraction situation, $2d \sin(\theta) = n\lambda$, where d is the distance between two planes, λ, θ are wavelength and angle to the normal, respectively, n is an integer, the waves are in phase, whereas in the destructive diffraction configuration, the waves have opposite phases

Analyzing the resulting pattern and the geometry of the experiment, in particular the angles of incidence and reflection, they found that the positions of the rings corresponded to angles such that two waves reflected from different atomic layers in the crystal were in phase, i.e., had their phases different by an integer multiple of 360°, as shown in Fig. 4.5a. The darkest areas corresponded to the situations when the reflected waves were out of phase, i.e., their phases were different by an odd integer multiple of 180°, thus canceling each other, as shown in Fig. 4.5b. By quantifying the positions of the rings, Davisson and Germer were able to confirm the de Broglie relation given in Eq. (4.3).

4.2 Elements of Quantum Mechanics

In this section, the essential quantum mechanics formalism and postulates and their mathematical treatment will be introduced. Their purpose will be to provide a general understanding of the behavior of electrons and energy band structures in solids and semiconductors, as discussed in subsequent sections.

4.2.1 Basic Formalism

The contradictions encountered when applying classical mechanics and electrodynamics to atomic processes, e.g., processes involving particles of small masses and at small separation from other particles, could only be resolved through a fundamental modification of basic physical concepts. The formalism which enabled the combining of the particle-like and wave-like properties of matter was created in 1920s by Heisenberg and Schrödinger and was called quantum mechanics, whose basic formalism and postulates we will now review.

1. The state of a system can be described by a definite (in general complex) mathematical function $\Psi(x, y, z, t)$, called the wavefunction of the system, which depends on the set of coordinates (x, y, z) of the quantum system and time t.
2. The wavefunction is a solution of the time-dependent Schrödinger equation (SE):

$$i\hbar \frac{\partial \Psi(x, y, z, t)}{\partial t} = H\Psi(x, y, z, t) \tag{4.4a}$$

where the operator H is called the "Hamiltonian" of the system and represents the total energy of the system in the form of mathematical operators. The sum of the kinetic and potential energy operator which make up the Hamiltonian are given by:

$$H = -\frac{\hbar^2}{2m}\nabla^2 + U(x, y, z, t) \tag{4.4b}$$

Note that the first term represents the kinetic energy of the particle and is a differential operator which acts on the wavefunction. The second term, the potential energy, keeps its classical form. One can think of the action of H on the wavefunction to be one of "measurement" of the total energy of the system.

3. The kinetic energy term is written in terms of the operator ∇^2 which is called the Laplacian and is defined in orthonormal coordinates in three dimensions by:

$$\nabla^2\Psi(x, y, z) = \frac{\partial^2\Psi(x, y, z)}{\partial x^2} + \frac{\partial^2\Psi(x, y, z)}{\partial y^2} + \frac{\partial^2\Psi(x, y, z)}{\partial z^2} \tag{4.4c}$$

$U(x,y,z,t)$ is the potential energy of the system considered, $\hbar$ is Planck's constant, and i is the complex number such that $i^2 = -1$.

The next principle of quantum (SE). Having solved the SE and found the wavefunctions, we have the following properties:

4. The probability that a physical measurement will result in values of the system coordinates in a volume $dxdydz$ around (x, y, z) at a time t is given by $|\Psi(x, y, z, t)|^2 dxdydz$.
5. The sum of the probabilities of all possible values of spatial coordinates of the system must be, by definition, equal to unity:

$$\int |\Psi(x, y, z, t)|^2 dxdydz = 1 \tag{4.5}$$

This equation is the normalization condition for the wavefunction.

4.2.2 General Properties of Wavefunctions and the Schrödinger Equation

The wavefunctions solution of the Schrödinger equation must satisfy a few properties, most of which are direct consequences of the mathematical formalism from which such functions are constructed.

The main property which will be used in the rest of the text is that the wavefunction and its first derivative must be finite, continuous, and single-valued in all space even if the system under consideration contains a surface or interface where the potential $U(x, y, z)$ has a finite discontinuity. But, in the case when the potential becomes infinite beyond this surface, the continuity of the derivative of the wavefunction does not hold anymore. This means that a particle cannot penetrate into a region where an infinite potential exists and therefore that its wavefunction becomes zero there.

Note *In classical physics the state of a system of particles is known when at any given time "t" we know all the spatial coordinates of the particles* $\{\boldsymbol{r}_i(t)\}$ *and all their momenta* $\{\boldsymbol{p}_i(t)\}$. *We can predict completely what is going to happen next when we know all the forces acting on the particles because we know the particles must obey Newton's laws. In principle we can therefore, with the knowledge of an initial state at time* $t = 0$, *compute and predict the exact trajectories that the particles will follow in space and know the momenta at each time. Now consider the difference to quantum, mechanics. In quantum mechanics, all we can possibly know about the system is its wavefunction* $\Psi(x, y, z, t)$ *which is obtainable by solving the Schrödinger equation (SE) given by* Eq. (4.4a). *Solving the SE means solving a differential equation with a given initial condition and only allowing the solutions which satisfy the differentiability and continuity conditions mentioned above. Now let us consider the next set of principles.*

Physical Observables and Measurement Introduction

The next principal of quantum mechanics is that for any physical variable, for example, position, momentum energy, etc., one can associate an operator f which "acts" on a wavefunction, i.e., differentiates, integrates, or simply multiplies it with another function. This operator represents a physical observable. It is like an act of measurement on the system. The mathematical operator in quantum mechanics which represents a physical observable is known and has been extracted by using a procedure which we do not need to discuss at this stage. The most important ones are listed in Table 4.1.

Table 4.1 Examples of common physical quantities and their associated operators

Physical quantity	Operator	Expectation value
x, y, z (coordinates)	x, y, z	$<x> = \int \Psi^* x \Psi dxdydz$
p_x, p_y, p_z (momentum)	$\frac{\hbar}{i}\frac{\partial}{\partial x}, \frac{\hbar}{i}\frac{\partial}{\partial y}, \frac{\hbar}{i}\frac{\partial}{\partial z}$	$< p_x >= \int \Psi^* \frac{\hbar}{i}\frac{\partial \Psi}{\partial x} dxdydz$
E (energy)	$i\hbar\frac{\partial}{\partial t}$	$< E >= \int \Psi^* i\hbar\frac{\partial \Psi}{\partial t} dxdydz$

Note *In order to proceed further, we are first going to consider situations in which the Hamiltonian of the system does not depend on time. This is the most common situation encountered in practice. It is the situation where we have a closed system and the total energy is conserved. Here we will learn how to extract further information from the solution of the Schrödinger equation and then proceed to some practical examples. At a later stage in* Chap. 10*, we will also consider time-dependent perturbations and return to consider the solutions of the time-dependent Schrödinger equation.*

4.2.3 The Time-Independent Schrödinger Equation

A particular and important situation for the Schrödinger equation is that for a closed system in a time-independent external field. Then, the right-hand side of Eq. (4.4a) does not contain time explicitly. In this case, the states of the system which are described by the wavefunction $\Psi(x, y, z, t)$ are called stationary states, and the total energy of the system is conserved (in time).

Let us now operate on the wavefunction with the energy operator expressed in terms of the time derivative. The action of $i\hbar\frac{\partial}{\partial t}$ on the wavefunction is like asking the question what is the energy of the system? Since we are assuming that energy is a constant, we find that the following relation must be satisfied:

$$i\hbar\frac{\partial\Psi(x, y, z, t)}{\partial t} = E\Psi(x, y, z, t) \tag{4.6}$$

But mathematically this means that the wavefunction $\Psi(x, y, z, t)$ must be a product of a function $\varphi(x, y, z)$ which solely depends on coordinates and an exponential function which depends only on time, such that:

$$\Psi(x, y, z, t) = \varphi(x, y, z)\exp\left(-\frac{i}{\hbar}Et\right) \tag{4.7}$$

This relation follows also directly from the theory of differential equations when the operator $\widehat{H}$ is independent of time. So inserting this expression into the Schrödinger equation in Eq. (4.4a) and eliminating the exponential term on both sides of the equation, we obtain:

$$-\frac{\hbar^2}{2m}\nabla^2\varphi(x, y, z) + U(x, y, z)\varphi(x, y, z) = E\phi(x, y, z) \tag{4.8}$$

which can be rewritten more concisely as:

$$\widehat{H}\,\phi_n = E_n\phi_n \tag{4.9}$$

This last expression is called the time-independent Schrödinger equation. The label "n" denotes the fact that the differential equation can have a spectrum of

solutions each corresponding to an allowed energy state of the system E_n with its corresponding wavefunction ϕ_n. When we know all the wavefunctions ϕ_n, we also know all the possible allowed energy levels of the system. Now we can say that when we measure the energy of the system, we must find the system in one of these eigenstates. We note that when the system is in a stationary state, the time dependence is only a phase factor that means it does not have any effect of the probability distribution. The spatial density is not changing or evolving in time; this is what one would expect.

In the time-independent picture, the total energy operator is $\widehat{H}$ which is also called the Hamiltonian of the system. Even though the total energy of the system does not change with time, the system can be in many different stationary energy states, called eigenstates ϕ_n. Each eigenstate has its own eigenvalue or energy E_n. The action of $\widehat{H}$ is again like an act of measurement of the energy state of the system which can produce, or one also sometimes say forces, the system to adopt an allowed energy state. Once the system has been prepared in an eigenstate φ_n with eigenvalue E_n, it will stay there forever unless it is disturbed by a perturbation which changes its total energy. So in quantum mechanics, and this is indeed fascinating, time may elapse, but the system stays in its eigenstate unless during this elapsing time, it also gets disturbed. So in quantum mechanics, one can say that when one considers a closed system in an eigenstate, time does not elapse for that eigenstate; it does not age, unless something happens which can change the state of the system.

We shall come back to this again later when we consider the "Heisenberg uncertainty principle."

Physical Observables and Measurement

What we did with the energy operator, we can now do with other physical observables. We first recall the following: for any physical variable, for example, position, momentum energy, etc., one can associate an operator f which "acts" on a wavefunction, i.e., differentiates, integrates, or simply multiplies it with another function. Like H the Hamiltonian for the total energy, this operator represents a physical observable. The most important ones are listed in Table 4.1. Every physical observable, or what is now operator, has a set of eigenfunctions and corresponding eigenvalues. Thus the operator $\widehat{f}$, for example (hat denotes that it is an operator), acting on the allowed wavefunction produces a number f_f or "eigenvalue." The eigenvalue corresponds to a possible value of the observable, when the wavefunction on which it operates is an "eigenstate" or also called "eigenfunction" of this operator, in other words if it satisfies the so-called eigenvalue equation:

$$\vec{f}\,\phi_f = f_f \phi_f \tag{4.10}$$

We say ϕ_f is an eigenfunction of $\widehat{f}$ and f_f, the corresponding eigenvalue. Eigenfunctions belonging to different eigenvalues are orthogonal; this means that their inner product is equal to 1 when the wavefunctions belong to the same eigenvalue, and 0 otherwise, or mathematically expressed:

$$\int dxdydz\phi_{f_1}^*(x,y,z)\phi_{f_2}(x,y,z) = \delta_{f_1 f_2} \tag{4.11}$$

Another property is that eigenfunctions of physical observables form a complete set. This means that they can be regarded as an infinite set of vectors which span the so-called Hilbert space such that any function χ can be represented as a linear combination of these eigenfunctions:

$$\chi(x,y,z) = \sum_f a_f \phi_f(x,y,z) \tag{4.12a}$$

Operators which are physical observables must have the property that the expectation value of the operator is a real number. Such operators are called Hermitian operators. For Hermitian operators it follows that the so-called matrix element of an operator f taken between two different eigenstates:

$$f_{ij} = \int d\ \vec{r}\ \Phi_i^* f \Phi_j \tag{4.12b}$$

satisfies the relation $f_{ij} = (f_{ji})^*$.

What we said about physical observables includes of course also the total energy operator $\widehat{H}$. The eigenstates of energy φ_n form a complete set and are orthogonal. Operators can have simultaneous eigenstates but not always. For example a free particle moving unhindered in space has eigenstates of momentum and energy which are the same functions. A particle moving in a box has energy eigenstates, but not momentum eigenstates. We shall see this later more clearly when we solve these problems explicitly.

Admixture of States

Let us imagine we have prepared the system in a stationary state or eigenstate. Then at some time later, it is disturbed by a perturbation which constitutes necessarily a time-dependent change, for example, a light pulse. The system no longer stays in its eigenstate but now goes into an admixture of eigenstates such as:

$$\Psi(x,y,z,t) = \sum_n a_n \varphi_n(x,y,z).\exp\left(-\frac{i}{\hbar}E_n t\right) \tag{4.13}$$

6. The system need not be in a pure state anymore or eigenstate of an observable; it can be in a superposition of such states. In which case if one undertook a measurement, one would find it in any one of the combination of such states as in Eq. (4.13). This leads us to the next definition.
7. The mean value or expectation value of a physical quantity represented by an operator f is what is measured experimentally, is denoted $<\widehat{f}>$, and is given by:

$$<\widehat{f}>= \int \Psi(x,y,z,t)^* f \Psi(x,y,z,t) dxdydz \tag{4.14}$$

where $\Psi(x,y,z,t)$ is the wavefunction of the system considered and $(\ldots)^*$ stands for complex conjugate. Thus if:

$$\Psi(x,y,z,t) = c_{f1}\Psi_{f1}(x,y,z,t) + c_{f2}\Psi_{f2}(x,y,z,t) \tag{4.15}$$

the expectation value $<\widehat{f}>$ is given by:

$$<\widehat{f}>= \left|c_{f1}\right|^2 f_1 + \left|c_{f2}\right|^2 f_2 \tag{4.16}$$

Examples of physical quantities, their associated operators, and expectation values are given in Table 4.1.

Thus one can interpret $|c_{f1}|^2$, $|c_{f2}|^2$ as the probability of finding the particle in the state f_1 and f_2, respectively, and indeed we must also have $|c_{f1}|^2 + |c_{f2}|^2 = 1$.

The problem one is confronted with after the system has been disturbed is to find the coefficients a_n of admixtures in the sum given by Eq. (4.13). This is done by solving the time-dependent Schrödinger equation in the presence of the disturbance and with given initial conditions as shown in Chap. 10.

4.2.4 The Heisenberg Uncertainty Principle

This very important principle says that one of the consequences of quantum mechanics is that one cannot have absolute knowledge of time and energy simultaneously and that this is not a theoretical abstraction but an experimental fact which is verified every day. One of the Heisenberg uncertainty principles (HUP) is therefore:

$$\Delta E \Delta t \sim \hbar \tag{4.17}$$

In other words if one knows the energy E to great accuracy, then one has a large uncertainty Δt, in time t and vice versa. Let us immediately apply this to a stationary state in energy, where clearly by definition, we know the energy level of the particle with absolute accuracy. The meaning of Eq. (4.17) is that in this case, we can say nothing about the time. Indeed the time dependence of the wavefunction as shown by Eq. (4.7) is only a phase, which has no consequence on the probability distribution, for example. Indeed, as we pointed out before, when in an eigenstate of energy, the particle does not evolve in time. It stays in that same energy level until it is disturbed by some perturbation. The perturbation makes the Hamiltonian change in time, and this allows the particle to admix with other eigenstates of different energy, which is the same thing as saying that the system can now evolve in time. The HUP also applies to momentum and space. If one knows the absolute position of a particle in space, then one cannot say anything about its momentum and vice versa, so we also have:

$$\Delta p_\mu \Delta r_\mu \sim \hbar \tag{4.18}$$

where p_μ; r_μ are x,y,z components of momentum space, respectively. We shall see later in more detail that one of the consequences of this rule is that a particle which is confined to a finite size box cannot have zero average momentum or kinetic energy.

4.2.5 The Dirac Notation

A convenient way of writing eigenstates and matrix elements or wavefunction overlap integrals was invented by Dirac. Here are some examples of the *Dirac notation* from which one can deduce the structure:

$$\begin{aligned} &\Psi_n\left(\vec{r}\right) \rightarrow |n\rangle \\ &\Psi_{n,k}\left(\vec{r}\right) \rightarrow |n,k\rangle \\ &\int_{-\infty}^{\infty} d\ \vec{r}\ \Psi_n^* \widehat{A} \Psi_m = \left\langle n \middle| \widehat{A} \middle| m \right\rangle \end{aligned} \tag{4.19}$$

In Eq. (4.19), the right-hand side $|m\rangle$ is called the "Ket vector." The left-hand side $\langle n|$ is called the "Bra vector." When one considers the expectation value of the product of two operators, one can expand over a complete set of eigenstates and write (m,l are arbitrary indices):

$$\left\langle n \middle| \widehat{A}\widehat{B} \middle| n \right\rangle = \int \Psi_n^* \widehat{A}\widehat{B}\, \Psi_n d\ \vec{r} = \sum_m \left\langle n \middle| \widehat{A} \middle| m \right\rangle \left\langle m \middle| \widehat{B} \middle| n \right\rangle \tag{4.20}$$

$$\sum_m |m\rangle\langle m| = 1 \tag{4.21}$$

The operator $P_m = |m\rangle\langle m|$ is called a projection operator because it projects a wavefunction onto a "part" or component of that wavefunction, i.e., it tells us how much of the state ϕ_m is in the wavefunction Ψ:

$$P_m \Psi = \langle m||\Psi\rangle|m\rangle = \phi_m\left(\vec{r}\right) \int d\ \vec{r}\ \phi_m^* \Psi \tag{4.22}$$

Assuming, as must be generally true, that the wavefunction must be in a linear combination of a complete set of basis states or eigenvectors:

$$\Psi = \sum_l a_l |l\rangle = \sum_l a_l \phi_l \tag{4.23}$$

Then it follows by substituting Eq. (4.23) into Eq. (4.22) using the orthogonality of the ϕ_m, and taking the sum, that the total projection reproduces the wavefunction again:

$$\sum_m P_m \Psi = \sum_m \langle m||\Psi\rangle|m\rangle = \sum_m \phi_m(\vec{r}) \int d\,\vec{r}\;\Psi^* \phi_m = \sum_m a_m \phi_m = \Psi \quad (4.24)$$

4.2.6 The Heisenberg Equation of Motion

There is a way of describing the relationship between operators and the time dynamics in quantum mechanics which is very elegant and most useful and called the equation of motion approach. To get there we first recall that the time-dependent Schrödinger equation can be written as:

$$i\hbar \frac{\partial \Psi(x,y,z,t)}{\partial t} = \widehat{H}\;\Psi(x,y,z,t) \quad (4.25)$$

And that the expectation value of an operator $\widehat{A}$ is by definition:

$$<\widehat{A}> = \int \Psi(x,y,z,t)^*\;\widehat{A}\;\Psi(x,y,z,t)dxdydz \quad (4.26)$$

Now consider how this expectation value changes with time, i.e., its time derivative:

$$\begin{aligned} \frac{d}{dt}<\widehat{A}> &= \int \frac{\partial}{\partial t}\Psi(x,y,z,t)^*\;\widehat{A}\;\Psi(x,y,z,t)dxdydz \\ &+ \int \Psi(x,y,z,t)^*\;\widehat{A}\;\frac{\partial \Psi}{\partial t}dxdydz \end{aligned} \quad (4.27)$$

The right-hand side is also in Dirac notation:

$$\frac{d<\widehat{A}>}{dt} = \frac{i}{\hbar}\langle\Psi|\;\widehat{H}\widehat{A} - \widehat{A}\widehat{H}\;|\Psi\rangle \quad (4.28)$$

$$\left[\widehat{H},\widehat{A}\right] = \widehat{H}\widehat{A} - \widehat{A}\widehat{H} \quad (4.29)$$

where the last line is, by definition, the commutator and written as:

$$\frac{d<\vec{A}>}{dt} = \frac{i}{\hbar}\left[\widehat{H},\widehat{A}\right] \quad (4.30)$$

which is called the equation of motion of the operator $\widehat{A}$ and is equivalent to the statement that if the operator commutes with the Hamiltonian, then it is a constant of the motion, which means it does not depend on time. The statement is that the

eigenstates of $\widehat{H}$ are also eigenstates of $\widehat{A}$. One interesting and very important result is, also, that one can define the time derivative of an operator by using the commutator with the Hamiltonian. For example, the velocity operator is indeed:

$$\frac{d\widehat{x}}{dt} = \frac{i}{\hbar}\left[\widehat{H}, \widehat{x}\right] \tag{4.31}$$

For a free particle Hamiltonian, one can check that the right-hand side of Eq. (4.31) is indeed $-i\hbar\frac{1}{m}\frac{\partial}{\partial x}$. Equation (4.31) is of great significance in theoretical physics and has no direct analogy in classical physics; it only has a formal analogue called the Poisson Bracket. In quantum mechanics, Eq. (4.31) is, in particular, a statement that the velocity operator depends on the structure of the Hamiltonian and is not always just given (x-direction) by the operator $-i\hbar\frac{1}{m}\frac{\partial}{\partial x}$. For example, when there are spin-orbit forces or magnetic fields involved, then the velocity operator involves also spin-dependent or magnetic field-dependent terms (we shall see this later in Chap. 5 and also again in Chap. 12). This has no simple analogue in classical physics. Processes or terms contained in $\widehat{H}$ which act on the position of the particle, and therefore do not commute with it, do not just give new energy levels but also give rise to new contributions to the definition of the velocity operator itself. Note that using Eq. (4.31), we can also define acceleration operators, for example, a_x where $a_x = \frac{i}{\hbar}[H, v_x]$. Equation (4.31) is the right and generally valid way of identifying the velocity operator in quantum mechanics. We shall see later that in magnetic field, the velocity operator is different from the free particle form given in Table 4.1; it has an extra term which depends on the field.

4.3 Discussion

As a first summary we note that whereas in classical mechanics one can in principle know energy, position, momentum, and time of a system simultaneously and with absolute accuracy, the same is not true in quantum mechanics. In quantum mechanics one can only at best know the wavefunctions which are the solutions of the Schrödinger equation (SE). Everything that can be known about the system must be deduced from the wavefunctions. This includes the probability distribution in space and the expectation value of the physical observables. Thus in quantum mechanics, the totality of solutions of the SE as we have seen form a complete set; in other words, the system can under all circumstances be found in a linear superposition of this complete set of eigenfunctions, each one belonging to an eigenvalue of energy. Similarly the thermal average of a physical observable $\widehat{A}$ is given by the generalized form of the Boltzmann distribution $<\widehat{A}>= \frac{\sum_n e^{-E_n/k_bT} A_{nn}}{\sum_n e^{-E_n/k_bT}}$ which involves the expectation values of the operator "$\widehat{A}$" A_{nn} over all the eigenstates of energy labeled by n. Unlike in classical mechanics where physical variables are defined irrespective

of the fact that they can be measured or not, in quantum mechanics, only measurable parameters are meaningful. These are the physical observables, and each one has its own operator representation. Measuring the value of a physical observable means calculating the expectation value of the operator, given that one has the wavefunction of the system. If the system is in a pure state, in other words in an eigenstate, then the outcome of this operation, or act of measurement, is the corresponding eigenvalue. In general, however, the system is in a superposition of eigenstates, and the outcome of the measurement is the weighted superposition as given by Eq. (4.13).

Note *The notion that in quantum mechanics, and thus in natural sciences, only measurable parameters are meaningful is also of deep philosophical significance, and the student should think about it carefully.*

4.4 Simple Quantum Mechanical Systems

4.4.1 Free Particle

The simplest example of solution of the Schrödinger equation is for a free particle of mass m and energy E, without external field and thus with a constant potential energy which can then be chosen to be zero $U(x, y, z) = 0$. For further simplicity, we can restrict the mathematical treatment to the one-dimensional time-independent Schrödinger equation. Eq. (4.8) can then be simplified to:

$$\frac{\hbar^2}{2m}\frac{d^2\Psi(x)}{dx^2} + E\Psi(x) = 0 \tag{4.32}$$

The solution of Eq. (4.32) which happens to be an eigenstate of both energy and momentum is:

$$\Psi(x) = Ae^{ikx} \tag{4.33}$$

where A is a constant and $k = \frac{2\pi}{\lambda}$ is the wavenumber. By applying the x-momentum operator on the right-hand side of Eq. (4.33), one can see that this state corresponds to a free particle state moving in the positive x-direction with momentum $\hbar k$. Replacing the expression of the wavefunctions into Eq. (4.32), one obtains:

$$-\frac{\hbar^2 k^2}{2m}\Psi(x) + E\Psi(x) = 0 \tag{4.34}$$

which has a nonzero solution for $\Psi(x)$ only if:

$$E = \frac{\hbar^2 k^2}{2m} = \frac{h^2 k^2}{8\pi^2 m} \tag{4.35}$$

or conversely:

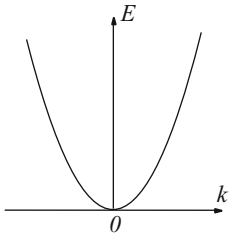


Fig. 4.6 The energy-momentum relationship for a free particle has a parabolic shape

$$k = \sqrt{\frac{2mE}{\hbar^2}} \tag{4.36}$$

and is plotted in Fig. 4.6. The particle momentum, as defined also by the expectation value, can be expressed in quantum mechanics as:

$$\langle p \rangle = \hbar k \tag{4.37}$$

The energy of the free particle depends therefore on its momentum as $E = \frac{\langle p \rangle^2}{2m}$, which is analogous to the case in classical mechanics. We can think of the system as very large and of size 2 L $\{-L, L)$, as L becomes infinite, so that the normalization constant A is given by $A = \sqrt{\frac{1}{2L}}$.

4.4.2 Degeneracy

The eigenstates with + and –k have the same energy; one says that the level k is twofold degenerate. Whenever an energy eigenstate has more than one quantum number which gives the same energy, one says that the level is degenerate.

4.4.3 Particle in a 1-D Box

Another simple and important illustration of quantum mechanics concepts can be obtained by considering a particle whose motion is confined in space. For simplicity, the analysis will be conducted in one dimension. It involves a particle of mass m and an energy E which evolves in a potential $U(x)$, shown in Fig. 4.7.

This potential can be mathematically expressed such that:

$$\begin{cases} U(x) = \infty & \text{for } x < 0 \text{ and } x > a \\ U(x) = 0 & \text{for } 0 < x < a \end{cases} \tag{4.38}$$

In such a potential, the properties of the wavefunctions and Schrödinger equation lead us to:

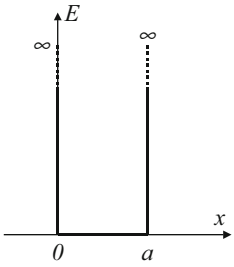


Fig. 4.7 Potential energy corresponding to the 1-D box

$$\begin{cases} \Psi(x) = 0 & \text{for } x < 0 \text{ and } x > a \\ \dfrac{\hbar^2}{2m}\dfrac{d^2\Psi(x)}{dx^2} + E\Psi(x) = 0 & \text{for } 0 < x < a \end{cases} \tag{4.39}$$

which means that the solution $\Psi(x)$ inside the box has the same expression as for the free particle in Eq. (4.33) and can be rewritten as the sum of *sin* and *cos* functions for simplification:

$$\Psi(x) = A\sin(kx) + B\cos(kx) \tag{4.40}$$

but with the boundary conditions:

$$\Psi(0) = \Psi(a) = 0 \tag{4.41}$$

Expressing these conditions using Eq. (4.40), we get, with $k = \frac{2\pi}{\lambda}$:

$$\begin{cases} B = 0 \\ A\sin(ka) = 0 \end{cases} \tag{4.42}$$

Since the wavefunction cannot be identically zero in the entire space, the following condition must be satisfied:

$$\sin(ka) = 0 \text{ or } k = k_n = n\frac{\pi}{a} \quad \text{where } n \text{ is an integer equal to } \pm 1, \pm 2, \ldots$$

Consequently, in contrast to the free particle case, not all values of the wavenumber k are allowed, but only discrete values are allowed. n can also be viewed as a quantum number of the system. Using Eq. (4.42), we can see that the energy of a particle in a 1-D box is also quantized:

$$E_n = n^2\frac{\hbar^2\pi^2}{2ma^2} \tag{4.43}$$

One can see that when $a \to \infty$, the spacing between the quantized energy levels tends toward zero and a quasi-continuous energy spectrum is achieved, as for a free particle. Nevertheless, the energy levels remain strictly discrete (this is why we talk

about a "quasi"-continuous energy spectrum). Combining Eq. (4.40) and Eq. (4.42), we can write the wavefunction as:

$$\Psi_n(x) = A \sin\left(\frac{n\pi x}{a}\right) \tag{4.44}$$

The value of A can be computed by substituting this expression into the normalization condition expressed in Eq. (4.5). One easily finds that:

$$A = \sqrt{\frac{2}{a}} \tag{4.45}$$

so that the complete analytical expression of the wavefunction solution of the infinite potential well problem is:

$$\Psi_n(x) = \sqrt{\frac{2}{a}} \sin\left(\frac{n\pi x}{a}\right) \tag{4.46}$$

These functions consist of standing waves as depicted in Fig. 4.8b. One can think of the particle in a 1-D box as bouncing on the walls of the box and the probability of finding a particle at x in the box is shown in Fig. 4.8c.

* Unlike the free particle case, the eigenstates of energy are no longer eigenstates of momentum. Operating with $-i\hbar\frac{\partial}{\partial x}$ on Eq. (4.46) does not give back the same function. Classically the particle is bouncing from the sides of the box and keeps on changing its momentum. The expectation value of the momentum can be evaluated as usual from Eq. (4.26) and can be verified to be zero.

Example

Q: Find the energy levels of an infinite quantum well that has a width of $a = 25$ Å.

A: The energy levels are given by the expression $E_n = n^2 \frac{\hbar^2\pi^2}{2m_0a^2}$, where m_0 is the free electron rest mass. This gives numerically:

$$\begin{aligned} E_n &= n^2 \frac{\left(1.05458 \times 10^{-34}\right)^2 \pi^2}{2\left(0.91095 \times 10^{-30}\right)\left(25 \times 10^{-10}\right)^2} \\ &= 9.63n^2 \times 10^{-21}\,\text{J} \\ &= 0.060n^2\,\text{eV} \end{aligned}$$

4.4.4 Particle in a Finite Potential Well

The infinite-potential analysis conducted previously corresponds to an unrealistic situation, and a finite potential well is more appropriate. Under these conditions, the potential in the Schrödinger equation is shown in Fig. 4.9 and mathematically expressed as:

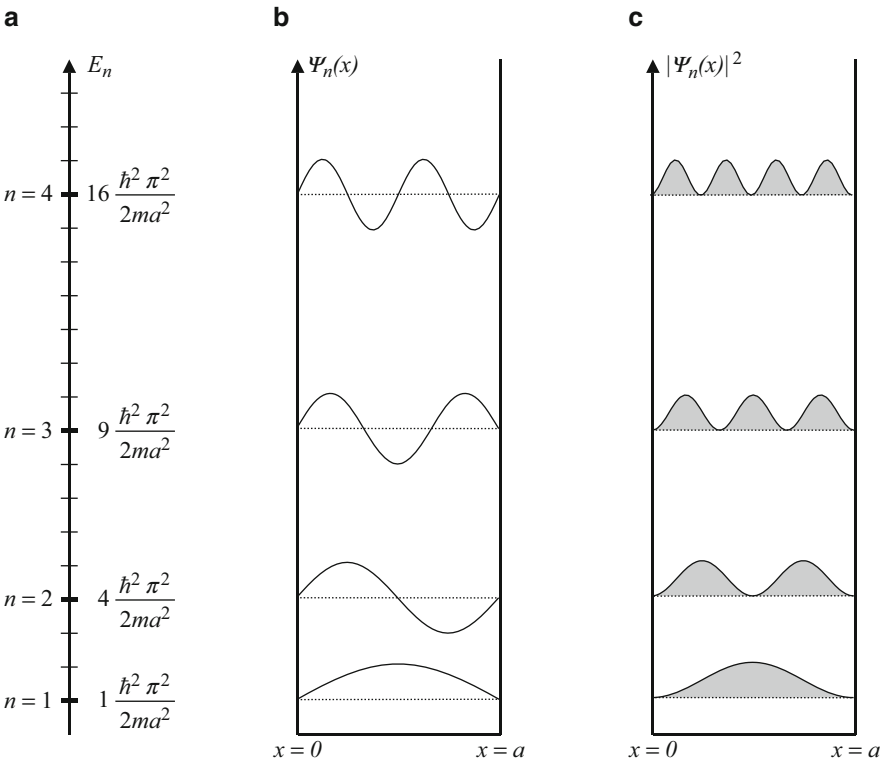


Fig. 4.8 (**a**) Energy levels, (**b**) wavefunctions $\Psi(x)$, and (**c**) $|\Psi(x)|^2$ which is proportional to the probability of finding a particle at a position x in a 1-D quantum box, for the first four allowed levels

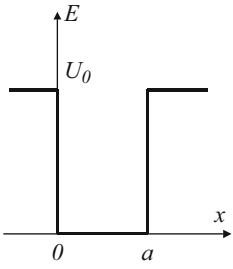


Fig. 4.9 Potential energy in a finite potential well

$$\begin{cases} U(x) = U_0 > 0 & \text{for } x < 0 \text{ and } x > a \\ U(x) = 0 & \text{for } 0 < x < a \end{cases} \tag{4.47}$$

In such a potential, the properties of the wavefunctions and Schrödinger equation lead us to:

$$\begin{cases} \dfrac{\hbar^2}{2m}\dfrac{d^2\Psi(x)}{dx^2} + (E - U_0)\Psi(x) = 0 & \text{for } x < 0 \text{ and } x > a \\ \dfrac{\hbar^2}{2m}\dfrac{d^2\Psi(x)}{dx^2} + E\Psi(x) = 0 & \text{for } 0 < x < a \end{cases} \tag{4.48}$$

We see that two distinct cases must be considered when solving this system of equations. The first one is when $0 < E < U_0$ and the other is when $U_0 < E$.

In the case of $0 < E < U_0$, Eq. (4.48) can be rewritten as:

$$\begin{cases} \dfrac{d^2\Psi(x)}{dx^2} - \alpha^2\Psi(x) = 0 & \text{for } x < 0 \text{ and } x > a \\ \dfrac{d^2\Psi(x)}{dx^2} + k^2\Psi(x) = 0 & \text{for } 0 < x < a \end{cases} \tag{4.49}$$

by defining:

$$\begin{cases} \alpha = \sqrt{\dfrac{2m(U_0 - E)}{\hbar^2}} \\ k = \sqrt{\dfrac{2mE}{\hbar^2}} \end{cases} \tag{4.50}$$

The general solution to Eq. (4.49) is then:

$$\begin{cases} \Psi_-(x) = A_- e^{\alpha x} + B_- e^{-\alpha x} & \text{for } x < 0 \\ \Psi_0(x) = A_0 \sin(kx) + B_0 \cos(kx) & \text{for } 0 < x < a \\ \Psi_+(x) = A_+ e^{\alpha x} + B_+ e^{-\alpha x} & \text{for } x > a \end{cases} \tag{4.51}$$

The boundary conditions include the finite nature of $\Psi(x)$ for $x \to \infty$ and $x \to -\infty$, the continuity of $\Psi(x)$, and its first derivative $\frac{d\Psi(x)}{dx}$ at points $x = 0$ and $x = a$, which can all be mathematically summarized as:

$$\begin{cases} \Psi_-(-\infty) = 0 & \Psi_+(+\infty) = 0 \\ \Psi_-(0) = \Psi_0(0) & \Psi_0(a) = \Psi_+(a) \\ \dfrac{d\Psi_-}{dx}(0) = \dfrac{d\Psi_0}{dx}(0) & \dfrac{d\Psi_0}{dx}(a) = \dfrac{d\Psi_+}{dx}(a) \end{cases} \tag{4.52}$$

Utilizing Eq. (4.52), we obtain:

$$\begin{cases} A_+ = B_- = 0 & \\ A_- = B_0 & A_0 \sin(ka) + B_0 \cos(ka) = B_+ e^{-\alpha a} \\ \alpha A_- = k A_0 & k A_0 \cos(ka) - k B_0 \sin(ka) = -\alpha B_+ e^{-\alpha a} \end{cases} \tag{4.53}$$

From these equations, we see that B_0 can be easily expressed in terms of A_0, and we thus obtain two equations involving only B_+:

$$\begin{cases} A_0\left[\sin(ka)+\dfrac{k}{\alpha}\cos(ka)\right]-B_+[e^{-\alpha a}]=0 \\ A_0\left[k\left(\cos(ka)-\dfrac{k}{\alpha}\sin(ka)\right)\right]+B_+[\alpha e^{-\alpha a}]=0 \end{cases} \tag{4.54}$$

A nonzero solution for A_0 and B_+, and thus a nonzero wavefunction, is possible only if:

$$\left(k^2-\alpha^2\right)\sin(ka)-2\alpha k\cos(ka)=0 \tag{4.55}$$

This condition can be rewritten into:

$$\tan(ka)=\frac{2\alpha k}{k^2-\alpha^2} \tag{4.56}$$

By introducing the constants:

$$\begin{cases} \alpha_0=\sqrt{\dfrac{2mU_0}{\hbar^2}} \\ \zeta=\dfrac{E}{U_0} \qquad (0<\zeta<1) \end{cases} \tag{4.57}$$

we can first rewrite Eq. (4.50) as:

$$\begin{cases} \alpha=\alpha_0\sqrt{1-\zeta} \\ k=\alpha_0\sqrt{\zeta} \end{cases} \tag{4.58}$$

and therefore:

$$\tan\left(a\alpha_0\sqrt{\zeta}\right)=\frac{2\sqrt{\zeta(1-\zeta)}}{2\zeta-1} \tag{4.59}$$

The only variable in Eq. (4.58) is ζ, and any value that satisfies leads to a value of E, k, and α and thus a wavefunction $\Psi(x)$ solution of the Schrödinger equation for the finite potential well problem in the case $0<E<U_0$.

Eq. (4.58) is easiest solved graphically. For example, Fig. 4.10 shows a plot of the two functions on either side of Eq. (4.58). The intersection points correspond to values of ζ which satisfy Eq. (4.58), and the number of intersection points is the number of bound states (i.e., wavefunction and energy level) in the finite potential well. In the example depicted in Fig. 4.10, there are two solutions. As the well potential U_0 increases, α_0 increases as defined by Eq. (4.57), and thus, a higher number of tangent function branches can be fitted for ζ between 0 and 1 (left-hand

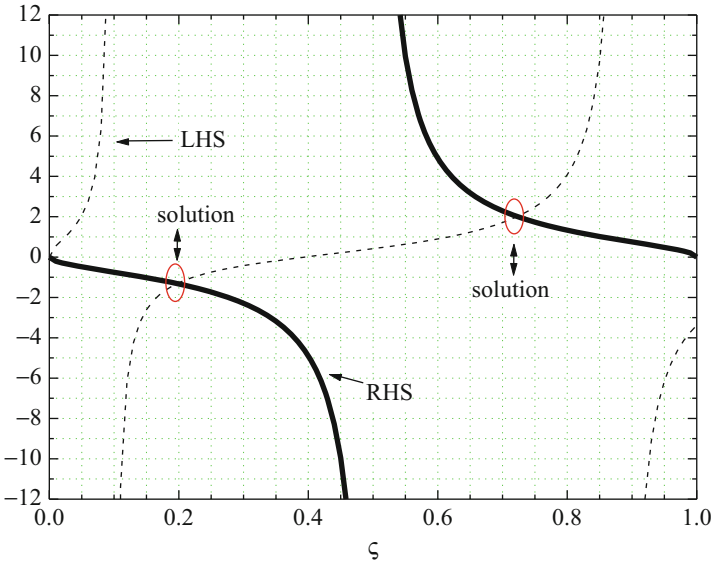


Fig. 4.10 Graphical representations of the functions on the left-hand side (LHS) and right-hand side (RHS) of Eq. (4.59), shown in dashed and solid lines. The intersections between these curves yield the solutions of the finite potential well problem

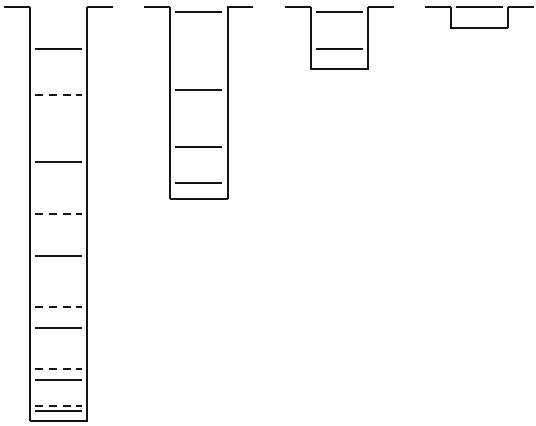


Fig. 4.11 Quantized energy levels in a finite potential well (solid lines) as a function of potential well depth. For comparison, the energy levels of the infinite well case are shown in dashed lines for the quantum well on the left

side of Eq. (4.59)). Consequently, the number of intersections solutions for ζ increases too, which means that there are more bound states in the well. This is schematically shown in Fig. 4.11. This can be understood intuitively because one can "fit" more bound states as the depth of the well increases.

Because there is only a discrete number of values for ζ, there is also a discrete number of energy values E, i.e., the energy levels are quantized similar to the infinite well potential case. In addition, the quantized values of energy here are found to be

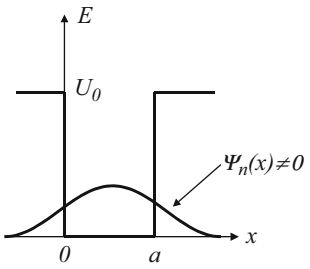


Fig. 4.12 Illustration of the tunneling effect in a finite potential well. The wavefunction is nonzero outside the potential well. This means that there exists a nonzero probability of presence for an electron outside the potential well is even when its energy E is lower than the potential barrier height U_0

lower than those in the infinite well potential case, as shown with the dashed lines in Fig. 4.11.

In addition to the quantization of energy levels, there is another important quantum concept illustrated by the finite potential well: the phenomenon of tunneling. Indeed, a nonzero wavefunction exists in the regions $x < 0$ and $x > a$, which means that the probability of finding a particle there is nonzero. In other words, even if a particle has an energy E lower than the potential barrier U_0, it has a nonzero probability of being found beyond the barrier. This is schematically shown in Fig. 4.12.

In the case of $E > U_0$, the solution of Eq. (4.49) can again as before be written as a sum of a cosine and sine term, for each of the regions defined by Eq. (4.51). Another more elegant way to represent the solution is as a sum of two plane waves, one going to the left and the other to the right. The two plane waves have different wavenumber k. The boundary conditions include the continuity of the wavefunction $\Psi(x)$ and its first derivative $\frac{d\Psi(x)}{dx}$ at points $x = 0$ and $x = a$. Along with the normalization condition expressed in Eq. (4.11), one can analytically determine the wavefunction. This analysis would lead to the same result as for a free particle, that is, there is a continuum of energy states $E > U_0$ allowed.

4.5 Discussion

In this chapter, we have shown the limitations of classical mechanics and the success of quantum mechanics. The basic concepts and formalism of quantum mechanics have been exposed, including the quantized nature of the electromagnetic field, the wave-particle duality, the probability of presence of a particle, the wavefunction, and the Schrödinger equation. Simple quantum mechanical systems have been analyzed to understand these novel major aspects associated with quantum mechanics have been discussed, including the quantization of energy levels and momenta and tunneling effects.

4.6 The Harmonic Oscillator

Recall from classical mechanics the motion of a particle moving in a one-dimensional force field which is linear in the displacement x with force constant K. Newton's law gives:

$$m\frac{d^2x}{dt^2} = -Kx \tag{4.60}$$

We solve this differential equation by noting that the solution is a simple sine function where:

$$x(t) = A\sin\omega_0 t \tag{4.61}$$

$$\omega_0^2 = \frac{K}{m} \tag{4.62}$$

The total energy is:

$$E = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}Kx^2 \tag{4.63}$$

and constitutes the classical Hamiltonian of the Harmonic oscillator problem. In quantum mechanics we can rewrite the Hamiltonian by making use of the definition of the momentum operator and keeping the potential energy as it is, to obtain:

$$H = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \frac{K}{2}x^2 \tag{4.64}$$

In order to obtain the energy levels and eigenfunctions of the harmonic oscillator, we have to solve the Schrödinger equation:

$$H\Psi_n = \left\{-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \frac{K}{2}x^2\right\}\Psi_n = E_n\Psi_n \tag{4.65}$$

For the mathematician this is a well-known differential equation which was solved long before the ideas of quantum mechanics were developed. We shall therefore here also treat it as a mathematical problem. More detailed developments can be found in specialized textbooks. The solution of Eq. (4.65) can be written with $\varsigma = \sqrt{\frac{m\omega_0}{\hbar}}x$:

$$\Psi_n = A_n H_n(\varsigma)\exp\left[-\frac{\varsigma^2}{2}\right] \tag{4.66}$$

Table 4.2 The first few wavefunctions and energies of the harmonic oscillator problem

n	E_n	Ψ_n
0	$\hbar\omega_0/2$	$A_0 \exp[-\varsigma^2/2]$
1	$3\hbar\omega_0/2$	$A_1 2\varsigma \exp[-\varsigma^2/2]$
2	$5\hbar\omega_0/2$	$A_2.(4\varsigma^2 - 2)\exp[-\varsigma^2/2]$
3	$7\hbar\omega_0/2$	$A_3(8\varsigma^3 - 12\varsigma)\exp[-\varsigma^2/2]$
4	$9\hbar\omega_0/2$	$A_4(16\varsigma^4 - 48\varsigma^2 + 12)\exp[-\varsigma^2/2]$
5	$11\hbar\omega_0/2$	$A_5(32\varsigma^5 - 160\varsigma^3 + 120\varsigma)\exp[-\varsigma^2/2]$
		$A_n = (2^n n!\sqrt{\pi})^{-1/2}$

$$E_n = \hbar\omega_0(n + 1/2) \tag{4.67}$$

where n is an integer which starts at $n = 0$ and A_n is a normalization constant defined by the requirement:

$$\int_{-\infty}^{\infty} dx \Psi_n^* \Psi_n = 1 \tag{4.68}$$

and where the H_n are the so-called Hermite polynomials which are tabulated. The new variable is related to the spatial variable x by:

$$\varsigma = \sqrt{\frac{m\omega_0}{\hbar}} x \tag{4.69}$$

The first few Hermite polynomials are given in Table 4.2 and plotted in Fig. 4.13. It is interesting to note that in the lowest energy, the ground state is not zero but a finite number given by $\hbar\omega_0/2$. This is called "the zero point vibrational energy." It is also a consequence of the Heisenberg uncertainty principle, because it is a manifestation of the fact that when a particle is confined in space by a potential, then its momentum and thus its energy can never be zero. This is one of the truly exciting features of quantum mechanics. The exact solution of the harmonic oscillator problem can be extended to the three-dimensional case without difficulty, provided the potential $V(x,y,z)$ is separable and a sum of the potentials in the three spatial directions:

$$V(x, y, z) = \frac{1}{2}\left\{K_x x^2 + K_y y^2 + K_z z^2\right\} \tag{4.70}$$

4.7 The Hydrogen Atom

As another most important example of the exact solution of a physical problem in quantum mechanics is the solution of the hydrogen atom problem, let us write down the total Hamiltonian of the electron and the proton nucleus with masses m_1 and m_2, respectively:

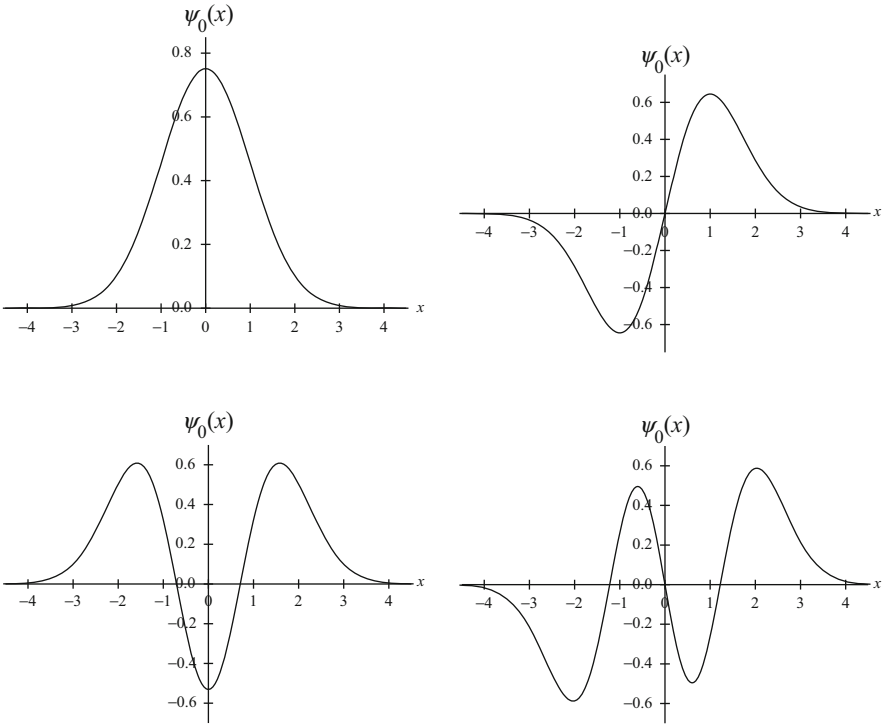


Fig. 4.13 The first few normalized wavefunctions of the harmonic oscillator with $n = 0$, $n = 1$, $n = 2$, and $n = 3$

$$H = \frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_2} - \frac{q^2}{4\pi\varepsilon_0 \left| \vec{r}_1 - \vec{r}_2 \right|} \tag{4.71}$$

r_1 and r_2 and p_1 and p_2 are the spatial and momentum coordinates of electron and proton, respectively. The proton mass is 1000 times heavier, and in any case it is useful to work in the relative coordinate system. Using the quantum mechanics operators, Eq. (4.71) becomes:

$$H = -\frac{\hbar^2}{2m_1}\nabla_1^2\Psi - \frac{\hbar^2}{2m_2}\nabla_2^2\Psi - \frac{q^2}{4\pi\varepsilon_0 \left| \vec{r}_1 - \vec{r}_2 \right|}\Psi = E\Psi. \tag{4.72}$$

where E is the total energy of the system. Now let us define the new center of mass variables:

$$\begin{aligned}\vec{R} &= [X, Y, Z] = \frac{m_1\vec{r}_1 + m_2\vec{r}_2}{m_1 + m_2} \\ \vec{r} &= \vec{r}_1 - \vec{r}_2\end{aligned} \tag{4.73}$$

Now we can write for the partial derivatives:

$$\nabla_1 = \nabla + \frac{m}{m_2}\nabla_R \tag{4.74}$$

$$\nabla_2 = -\nabla + \frac{m}{m_1}\nabla_R \tag{4.75}$$

$$\nabla_R = \left[\frac{\partial}{\partial X}, \frac{\partial}{\partial Y}, \frac{\partial}{\partial Z}\right] \tag{4.76}$$

$$\nabla = \left[\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right] \tag{4.77}$$

$$m = \frac{m_1 m_2}{m_1 + m_2} \tag{4.78}$$

Substituting back in terms of the new coordinates into the original Schrödinger equation, we have:

$$-\frac{\hbar^2}{2(m_1 + m_2)}\nabla_R^2\Psi - \frac{\hbar^2}{2m}\nabla^2\Psi - \frac{q^2}{4\pi\varepsilon_0|r|}\Psi = E\Psi \tag{4.79}$$

Now in this form, we see that the differential equation is separable in terms of the relative electron-nucleus, and center of mass motion of the atom, so that the total wavefunction can be written mathematically as a product:

$$\Psi\left(\vec{R}, \vec{r}\right) = \Phi\left(\vec{R}\right)\psi\left(\vec{r}\right) \tag{4.80}$$

Substitute back into Eq. (4.79) and rewrite the total equation in terms of two separate ones:

$$-\frac{\hbar^2}{2(m_1 + m_2)}\nabla_R^2\Phi\left(\vec{R}\right) = E_c\Phi\left(\vec{R}\right) \tag{4.81}$$

$$\begin{aligned}&-\frac{\hbar^2}{2(m)}\nabla^2\psi\left(\vec{r}\right) - \frac{q^2}{4\pi\varepsilon\varepsilon_0 r}\Psi\left(\vec{r}\right) = E_r\psi\left(\vec{r}\right) \\ &E = E_c + E_r\end{aligned} \tag{4.82}$$

The center mass motion is free and can therefore be solved immediately:

$$\Phi(\bar{R}) = C\exp\left[(i/\hbar)\ \vec{P}\cdot\vec{R}\right] \tag{4.83}$$

where $\boldsymbol{P}$ is the center of mass momentum, C the normalization constant, and the magnitude of $\boldsymbol{P}$ is related to the center of mass energy by the equation:

$$\left|\vec{P}\right| = \sqrt{2(m_1+m_2)E_c} \tag{4.84}$$

Now let us consider the relative motion of the electron around the nucleus. It is convenient to work in atomic energy units in which the energy is measured in multiples of the ionization energy of hydrogen which is the Rydberg unit $R = \frac{mq^4}{2(4\pi\varepsilon_0)^2\hbar^2}$ and we measure the coordinates, i.e., lengths, in units of the Bohr radius $a_B = \frac{\hbar^2(4\pi\varepsilon_0)}{mq^2}$. When working in terms of these units, we can put $\hbar = 1$; $q^2 = 2$; $m = /12$ in Eq. (4.82) to find the dimensionless form (we have dropped the index r on the energy for convenience):

$$\nabla^2\psi(\vec{r}) + \left(E + \frac{2}{r}\right\}\Psi(\vec{r}) = 0 \tag{4.85}$$

Equation (4.85) has the special feature that it is now an equation involving a particle moving in a spherically symmetric potential. We can now solve it as a mathematical problem by exploiting the spherical symmetry of the problem. In doing so we will naturally encounter the concept of angular momentum.

4.7.1 Motion in a Spherically Symmetric Potential

Given the symmetry of the problem, it is convenient to work with spherical polar coordinates. The differential Eq. (4.85) can be rewritten using:

$$\begin{aligned} x &= r\sin\theta\cos\phi \\ y &= r\sin\theta\sin\phi \\ z &= r\cos\theta \end{aligned} \tag{4.86}$$

and thus Eq. (4.85) becomes:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2}\left\{\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2} + \left(E - \frac{2}{r}\right)\right\}\psi(r,\theta,\phi) = 0 \tag{4.87}$$

Again, this equation has a separable structure, in which the angular part and the radial part can be considered to vary independently so that:

$$\psi = R(r)Y(\theta, \phi) \tag{4.88}$$

Substituting Eq. (4.88) back into Eq. (4.87) allows us to rearrange the equation into the form:

$$\frac{1}{R}\left\{\left[\frac{d}{dr}\left(r^2\frac{\partial R}{\partial r}\right) + \left[E - \frac{2}{r}\right]r^2R(r)\right\} = -\frac{1}{Y}\left\{\frac{1}{\sin\theta\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 Y}{\partial\phi^2}\right\} \tag{4.89}$$

The left-hand side (LHS) of this equation only depends on r, the right-hand side (RHS), only on the angles. The equation can be satisfied if each side of it is equal to the same constant C_0, so that:

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \left[E - \frac{2}{r} - \frac{C_0}{r^2}\right]R = 0 \tag{4.90}$$

$$\left\{\frac{1}{\sin\theta\partial\theta}\left(\sin\theta\frac{\partial Y}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 Y}{\partial\phi^2}\right\} = -C_0Y \tag{4.91}$$

The angular equation for Y can be further separated by writing:

$$Y(\theta, \phi) = P(\theta)\Phi(\phi) \tag{4.92}$$

and thus by substituting into Eq. (4.90) and Eq. (4.91), we have:

$$\frac{1}{P}\left[\sin\theta\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right)\right] + C_0\sin^2\theta = -\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2} = m^2 \tag{4.93}$$

In anticipation of the mathematical structure, we have introduced a separation constant which we have called m^2, and this allows us to rewrite the right-hand side of Eq. (4.93) as:

$$\frac{d^2\Phi}{d\phi^2} + m^2\Phi = 0 \tag{4.94}$$

which has a simple solution of the form (normalization will be handled later):

$$\Phi = \exp[\pm im\phi] \tag{4.95}$$

The LHS of Eq. (4.93) can be conveniently written in terms of a new variable $\mu = \cos\theta$ giving us:

$$\frac{d}{d\mu}\left\{(1-\mu^2)\frac{dP}{d\mu}+\left(C_0-\frac{m^2}{1-\mu^2}\right)P=0\right\} \tag{4.96}$$

This equation is well known to mathematicians, and, indeed, it is one of those fortunate facts that they had looked at this type of equation long before they were of relevance to quantum mechanics and studied them in detail. The scientific community would be very much worse off if these solutions had not be found before, and we had to compute the results numerically. In any case, as it happens, this equation is known as Legendre's equation, and it was discovered that it only had bounded and differentiable solutions if the constant:

$$C_0 = l(l+1) \tag{4.97}$$

where l is a positive integer and the values of m are also restricted to the range $\{-l, -l+1, \ldots .l, l+1\}$. Combining Eq. (4.97) and Eq. (4.96), we can now write down the complete solution of the angular part of the Schrödinger equation as:

$$Y_l^m(\theta,\phi) = A_l^m e^{im\phi} P_l^m(\cos\theta) \tag{4.98}$$

The P_l^m are called the Legendre polynomials, they are tabulated as special functions, and A_l^m are normalization constants which we will now give as the final form:

$$Y_l^m(\theta,\phi) = (-)^m \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}\, e^{im\phi} P_l^m(\cos\theta) \tag{4.99}$$

In order to complete the solution of the hydrogen atom, we still need the solution to the radial part $R(r)$. But before doing that, let us first understand the significance of the angular part.

4.7.2 Angular Momentum

When a system is rotationally invariant, we expect on grounds of symmetry theory, and classical physics, that the particle moving in such a spherically symmetric field should have a well-defined angular momentum. So we ask: What is the angular momentum of an electron moving in the orbital of a hydrogen atom? In order to answer this question, we first have to find the angular momentum operator $\widehat{L}$ in quantum mechanics. We do this as with other operators, we use the classical correspondence principle which says that if:

$$\widehat{\vec{L}} = \vec{r} \times \vec{p} = -\vec{p} \times \vec{r} \tag{4.100}$$

is the classical angular momentum, then the quantum mechanical operator is simply given by replacing the r and p by the corresponding values based on quantum mechanics. Thus, for example:

$$L_z = xp_y - yp_x = x\left(\frac{\hbar}{i}\frac{\partial}{\partial y}\right) - y\left(\frac{\hbar}{i}\frac{\partial}{\partial x}\right) \tag{4.101}$$

Then we note that the angular momentum operators in all three directions commute with the Hamiltonian of the hydrogen atom, i.e., the operators satisfy:

$$[H, L_x] = [H, L_y] = [H, L_z] = 0 \tag{4.102}$$

Equation (4.102) implies that eigenfunctions of $\widehat{H}$ are simultaneous eigenfunctions of $\widehat{L}$. An electron which is in an eigenstate of the Hydrogen atom is also in an eigenstate of angular momentum. In other words, the particle has both a well-defined energy and angular momentum. So we ask the question what is the angular momentum of the electron in the state $Y_l^m(\cos\theta)$ since we know this to be an eigenfunction? To answer this question, we make a measurement or apply the operator $\widehat{L}$ on the wavefunction. It is convenient and easier to work with the square of the angular momentum rather than the angular momentum itself. So we consider the operator:

$$L^2 = L_x^2 + L_y^2 + L_z^2 \tag{4.103}$$

and note that:

$$[L^2, L_x] = [L^2, L_y] = [L^2, L_z] = 0 \tag{4.104}$$

Also we have that:

$$[H, L^2] = 0 \tag{4.105}$$

It now follows that the energy eigenstates are simultaneous eigenstates of both L_z and L^2, but from vector algebra, it follows also that:

$$L^2\Psi = -\hbar^2\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\Psi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2\Psi}{\partial\theta}\right] \tag{4.106}$$

and this is exactly the same differential form as Eq. (4.91). So from Eq. (4.91) and Eq. (4.99), it follows that a measurement of the squared angular momentum on the state Y_l^m must give the output $\hbar^2 l(l+1)$ or in other words:

$$L^2 Y_l^m = -\hbar^2\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial Y_l^m}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2 Y_l^m}{\partial\theta}\right] = l(l+1)\hbar^2 Y_l^m \tag{4.107}$$

with the amplitude $\left|\vec{L}\right| = \hbar\sqrt{l(l+1)}$. Also a measurement of the z-component gives:

$$L_z Y_l^m = \frac{\hbar}{i}\frac{\partial}{\partial \phi} Y_l^m = m\hbar Y_l^m \tag{4.108}$$

so that a measurement of the projection of the angular momentum in the z-direction of the state $\left(Y_l^m\right)$ gives an eigenvalue $m\hbar$.

Now we understand the physical significance of the solutions that we derived in Sect. 4.7.1, and we also note the generality of the result. The eigenfunctions of angular momentum are the functions $\left(Y_l^m\right)$ and this is true in general. It happens to be true for the hydrogen atom too because the potential is spherically symmetric. So in any state with spherical symmetry, angular momentum is well defined, and the $\left(Y_l^m\right)$ constitutes the angular part of the wavefunction.

In this section we have tackled the solution of the hydrogen atom problem in quantum mechanics. We showed that the wavefunction can be written as a product of an angular and radial part. Now we can turn to studying the radial part $R(r)$ for this particular Coulomb potential.

4.7.3 The Radial Wavefunction of the Hydrogen Atom

Returning to face the solution of the radial part, we first note that a more convenient way of writing this equation is to transform:

$$u(r) = rR(r) \tag{4.109}$$

$$\frac{d^2u(r)}{dr^2} + \left[E + \frac{2}{r} - \frac{l(l+1)}{r^2}\right]u(r) = 0 \tag{4.110}$$

Now we note that differential equations involve the angular momentum integers l, so it follows that the eigenstates $u(r)$ must also have the label l, $u = u_l$, but there is also an energy variable E. So what happens to the energy E? Are all values allowed? It turns out not surprisingly that the answer is no! Again the mathematicians saw that long before the physicist used these solutions for the hydrogen atom. Mathematicians found that in order to have bounded and differentiable solutions, only discrete values of E were allowed. These carry the label n, and we have E_n as eigenvalues of energy and thus $u_{nl}(r)$, as eigenstates. The solution of this class of differential equation is a nontrivial exercise in mathematics, so we will only give the final answer here. The normalized solutions can be written as:

$$u_{n,l} = \sqrt{\frac{2r}{n^3}}\Lambda_{n-l-1}^{2l+1}\left(\frac{2r}{n}\right) \tag{4.111}$$

so that the complete solution is:

$$\Psi_{n,l,m} = \frac{1}{r} u_{n,l} Y_l^m(\cos\theta) \tag{4.112}$$

with energy eigenvalues:

$$E_n = -\frac{mq^4}{2n^2(4\pi\varepsilon_0)^2\hbar^2} \tag{4.113}$$

The functions Λ are related to the so-called Laguerre functions where with the variable $t = 2r/n$, we have:

$$\Lambda_k^\alpha(t) = \left[\Gamma(\alpha+1)\binom{k+\alpha}{k}\right]^{-1/2} e^{-t^2/2} t^{\alpha/2} L_k^\alpha(t) \tag{4.114}$$

where now L_k^α is a solution of Laguerre's differential equation:

$$t\frac{d^2L_k^\alpha}{dt^2} + (\alpha+1-t)\frac{dL_k^\alpha}{dt} + kL_k^\alpha = 0 \tag{4.115}$$

where α is a constant and Γ is the Gauss gamma-function with:

$$\binom{k+\alpha}{k} = \frac{\Gamma(k+\alpha+1)}{\Gamma(k+1)\Gamma(\alpha+1)} \tag{4.116}$$

The first few radial functions in atomic units are given by:

$$u_{10} = 2re^{-r} \tag{4.117}$$

$$u_{20} = \frac{1}{\sqrt{8}} e^{-r/2} r(2-r) \tag{4.118}$$

$$u_{21} = \frac{1}{\sqrt{24}} e^{-r/2} r^2 \tag{4.119}$$

$$u_{30} = \frac{2}{81\sqrt{3}} e^{-r/3} r\left(27 - 18r + 2r^2\right) \tag{4.120}$$

In real units the ground-state radial part is $R(r)_{10} = u_{10}/r = \frac{1}{\sqrt{8\pi}}\left(\frac{2}{a_B}\right)^{3/2} e^{-r/a_B}$ where a_B is the Bohr radius. The first few angular functions are (Fig. 4.14):

$$Y_0^0 = \frac{1}{\sqrt{4\pi}} \tag{4.121}$$

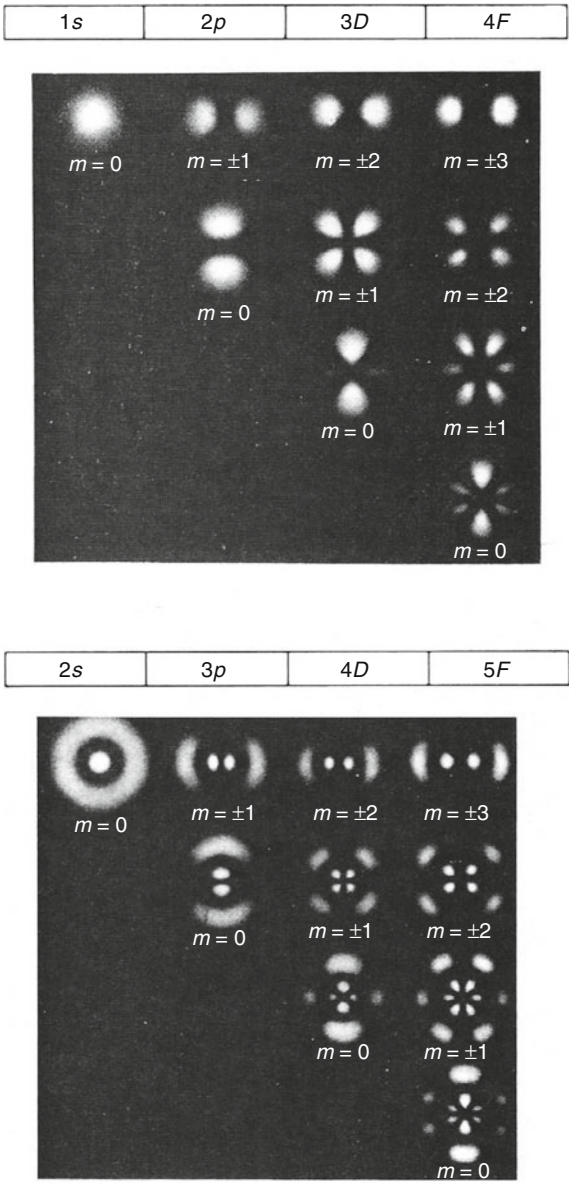


Fig. 4.14 Illustrates the particle density in the first few levels of the hydrogen atom. These are the 2S ($l = 0$), 2P ($l = 1$) and 3D($l = 2$) and 4F ($l = 3$) orbitals to the with the m, projections along the z-axis

$$Y_1^0 = \sqrt{\frac{3}{4\pi}}\cos\theta \tag{4.122}$$

$$Y_2^0 = \sqrt{\frac{5}{16\pi}}\left(3\cos^2\theta - 1\right) \tag{4.123}$$

$$Y_1^1 = -\sqrt{\frac{3}{8\pi}}\sin\theta\exp i\varphi \tag{4.124}$$

$$Y_2^1 = -\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta\exp[i\varphi] \tag{4.125}$$

4.7.4 The Unbound States

The above solutions encompass the region $E < 0$, i.e., where the electron is bound to the nucleus. One can also solve for the wavefunctions and energies of eigenstates which are free particle-like (only energies) in the region $E > 0$. The reader is referred to the book by *L*. Chuang in the references for a complete analytical description.

4.7.5 The Two-Dimensional Hydrogen Atom

Another interesting limit is the two-dimensional hydrogen atom which, even though it strictly speaking does not exist, is almost realizable using quantum well technology (see Chap. 15). One can, with atom-by-atom deposition techniques, molecular beam epitaxy (MBE), for example, place a hydrogenic atom in a thin atomic layer sandwiched between barrier layers so that the electronic motion is free in the plane and highly confined in the direction perpendicular to the plane. The analytical solution for the two-dimensional hydrogen atom problem is known mathematically, and the bound states are given by (Chuang 1995):

$$E_n = -\frac{R_y}{(n-1/2)^2}; n = 1, 2, 3\ldots. \tag{4.126}$$

$$R_y = \text{Rydberg} = \frac{mq^4}{2(4\pi\varepsilon_0)^2\hbar^2} \tag{4.127}$$

Comparing with the 3D solution Eq. (4.113), it is interesting to note that the binding energy is stronger in 2D than in 3D, a factor of 4 for the ground state. The wavefunctions only depend on one angle and have the simpler structure:

$$\Psi_{nm}\left(\vec{r}\right) = R_{nm}(r)\frac{e^{im\varphi}}{\sqrt{2\pi}} \tag{4.128}$$

Again we refer the reader to the book by *L.* Chuang in the references for the complete analytical formulae of the radial part and a discussion of the unbound solutions.

4.7.6 The Electron Spin

It was pointed out at the beginning of this chapter that in order to explain the structure of the atom, very early on after the discovery of quantum physics, W. Pauli introduced the concept of the electron spin. He observed that if he assumed that an electron had an extra quantum number which he called spin and if he assumed that the spin is like an angular momentum and can have two values "up or down" with values $\pm\frac{1}{2}\hbar$ and, further, assumed that two electrons cannot be in exactly the same eigenstate, then he could account for the so-called Aufbau principle i.e., explain the structure of the atoms with quantum mechanics.

So we have learned already that the electron must have a quantum number called spin, which is like an angular momentum and can have two values of its projection in the z-direction $S_z = \pm 1/2$. In other words if we make a measurement of the electron spin in a given direction which we call z, then we will find the values $S_z = \pm 1/2$ with equal probability. Once the electron has been prepared in a given spin state, it will stay in it unless disturbed. The electron spin is like an angular momentum in the sense that it carries also a magnetic moment as a classical rotating charge would do in principle. The magnitude of the magnetic moment was postulated and then measured by Stern and Gerlach to be $\mu_B = \frac{q\hbar}{m_0}$ (m_0 is the rest mass) which is called the Bohr magneton. The projection along the z-axis is $m_z = \pm 1/2\mu_B$.

Now the next question that arises is: where does the spin come from? Is it really due to a kind of zero point rotation of the electron in space, a rotating charge as one would think classically? Zero point meaning that according to Heisenberg's uncertainty principle, and as shown explicitly for the harmonic oscillator, when a degree of freedom is allowed, it has to have a "minimum value" associated with it. Otherwise one would know the position of a point on the surface with certainty. However classical rotation, it turns out, cannot be the reason for the electron spin and magnetic moment, because if one calculates the speed at which the charge would have to rotate to give this value of magnetic moment, one would find that the speed of rotation would be greater than the speed of light and therefore contrary to the rules of special relativity.

4.8 Relativity and Quantum Mechanics

The explanation for the electron spin came much later and was given by Paul A. M. Dirac in 1927. Dirac set himself the task of including special relativity into quantum mechanics and used the classical correspondence principle. Using Einstein's formula, gives the energy of the particle as:

$$E = c\sqrt{p^2 + m_0^2 c^2} \tag{4.129}$$

where m_0 is the rest mass and c the velocity of light. In this form, substituting in the Schrödinger momentum, operators would lead to a Hamiltonian which depends on the square root of differential operators. This is awkward to handle and is not evidently Lorentz invariant. Invariance under a Lorentz transformation is essential for special relativity to be satisfied. Of course one can develop the square root using the Taylor expansion assuming that the kinetic energy is small compared to the rest mass. The first two terms then together give back the usual result with a constant rest mass as first term. One can also write down the entire series, as an expansion in p^2:

$$H = m_0 c^2 + \frac{p^2}{2m_0} - \frac{1}{8}\frac{p^4}{m_0^3 c^2} + \ldots \tag{4.130}$$

Then using the quantum mechanical operators, assume that the eigenfunctions are plane waves as free particles and then every term in the series is a simple number with $p^2 \rightarrow (\hbar k)^2$. Resuming the series then gives the energy levels:

$$E_k = c\left[(\hbar k)^2 + m_0^2 c^2\right]^{1/2} \tag{4.131}$$

with a group velocity:

$$v_k = \frac{1}{\hbar}\frac{\partial E_k}{\partial k} = \frac{(\hbar k)c}{\left[m_0^2 c^2 + (\hbar k)^2\right]^{1/2}} \tag{4.132}$$

which saturates at the speed of light when the momentum becomes infinite. But this solution is not complete, and in order to ensure Lorentz invariance, Klein-Gordon and Dirac noted that one should consider the square of the operator and then replace the momentum and energy using the corresponding Schrödinger operators to find:

$$\left\{\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right\}\Psi = \left(\frac{m_0 c^2}{\hbar^2}\right)\Psi \tag{4.133}$$

This equation is known as the Klein-Gordon equation. It is second order in the time derivative as is Maxwell's equation (ME) and indeed is a wave equation as ME, when the particle has zero rest mass. The important observation is that this equation as is ME is relativistically invariant, i.e., it satisfies the Lorentz transformation

symmetry. This by the way also shows that special relativity is naturally true for electromagnetism and is thus an experimental law. Going back to Eq. (4.133), we note that we had to take the square or complex conjugate to arrive at Eq. (4.133), and therefore it will have more solutions than we may need. However it will certainly have all the solutions that we need. The energy operator is now as one can see quadratic in structure and no longer linear as in the nonrelativistic Schrödinger theory. The Klein-Gordon equation for free particles will also have the plane wave solutions discussed in the square expansion form, with the same energy-momentum relations, as can be easily verified by substituting:

$$\Psi\left(\vec{r},t\right) = A\exp\left\{i\,\vec{k}\cdot\vec{r}\,-iEt/\hbar\right) \tag{4.134}$$

Dirac's brilliant observation, which is the starting point of all modern quantum field theories and elementary particle descriptions until today, was to note that maybe one could go back to a linear form in terms of time and write this equation as the product of two linear differential equations. Let us write (Dirac PAM 1967):

$$\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} = \left(A\frac{\partial}{\partial x} + B\frac{\partial}{\partial y} + C\frac{\partial}{\partial z} + \frac{i}{c}D\frac{\partial}{\partial t}\right)\left(A\frac{\partial}{\partial x} + B\frac{\partial}{\partial y} + C\frac{\partial}{\partial z} + \frac{i}{c}D\frac{\partial}{\partial t}\right) \tag{4.135}$$

In order for this differential operator relation to be satisfied, we need:

$$\begin{aligned} AB + BA &= 0 \\ AC + CA &= 0 \\ AD + DA &= 0 \\ BC + CB &= 0 \\ BD + DB &= 0 \\ CD + DC &= 0 \end{aligned} \tag{4.136}$$

$$A^2 = B^2 = C^2 = D^2 = 1 \tag{4.137}$$

Dirac observed that this decomposition is possible provided that we do not look at these $\{A,B,C,D\}$ quantities as simple numbers but as matrices. He then solved the matrix problem to find:

$$(A,B,C) = i\beta\alpha_k \tag{4.138}$$

$$D = \beta \tag{4.139}$$

$$\beta = \begin{pmatrix} I, 0 \\ 0, -I \end{pmatrix} \tag{4.140}$$

$$\alpha_k = \begin{pmatrix} 0, \sigma_k \\ \sigma_k, 0 \end{pmatrix} \tag{4.141}$$

$$\sigma_k = \begin{pmatrix} 0, 1 \\ 1, 0 \end{pmatrix}; \begin{pmatrix} 0, -i \\ i, 0 \end{pmatrix}; \begin{pmatrix} 1, 0 \\ 0, -1 \end{pmatrix} \tag{4.142}$$

where I is the unit matrix. Note that A, B, and C are 4×4 matrices. With the above matrix form, the Klein-Gordon equation can be rewritten:

$$\left(A\frac{\partial}{\partial x} + B\frac{\partial}{\partial y} + C\frac{\partial}{\partial z} + \frac{i}{c}D\frac{\partial}{\partial t}\right)\left(A\frac{\partial}{\partial x} + B\frac{\partial}{\partial y} + C\frac{\partial}{\partial z} + \frac{i}{c}D\frac{\partial}{\partial t}\right) - \frac{m_0^2c^2}{\hbar^2}I = 0 \tag{4.143}$$

The linearized form is thus:

$$\left(A\frac{\partial}{\partial x} + B\frac{\partial}{\partial y} + C\frac{\partial}{\partial z} + \frac{i}{c}D\frac{\partial}{\partial t}\right) - \frac{m_0c}{\hbar}I = 0 \tag{4.144}$$

One can also rewrite these equations as the Dirac equation pair:

$$\begin{pmatrix} m_0c^2, c\,\vec{\sigma}\cdot\vec{p} \\ c\,\vec{\sigma}\cdot\vec{p}\,, -m_0c^2 \end{pmatrix}\begin{pmatrix} \phi^+ \\ \phi^- \end{pmatrix} = i\hbar\frac{\partial}{\partial t}\begin{pmatrix} \phi^+ \\ \phi^- \end{pmatrix}. \tag{4.145}$$

Note that these are two distinct coupled 2×2 differential equations. One has positive, the other negative energy solutions. Each component is itself a 2×2 matrix. The $\varphi+$ component is connected to the $\varphi-$ component by a relativistic coupling. In the nonrelativistic limit, the two are not coupled anymore, but the 2×2 matrix structure of each remains. The 2×2 matrix form implies that the particle, apart from its usual spatial degrees of freedom, must also have acquired an additional two-valued degree of freedom. This new degree of freedom is exactly the spin which had been earlier postulated by Pauli to also have exactly this matrix representation. What it means is that the wavefunction of a particle which satisfies the linear Dirac equation has two components, a component with an internal degree of freedom which can be called spin up and the other one spin down. This internal degree of freedom turns out to have the same properties as an angular momentum with the two possible values $\pm\frac{1}{2}\hbar$ as discussed earlier.

This is a remarkable achievement indeed and shows that the symmetry associated with special relativity in quantum mechanics has important consequences on the structure of the basis states of space, on the "fabric of space time," (Wilczek 2006) making the wavefunction 4 component with an extra two-valued degree of freedom. In relativity, space and time are connected, but the time derivative still measures the energy. So to be in an eigenstate of energy, has implications for the spatial coordinates. But this is not all; Dirac's equation implies that along with positive energy solutions, there are also negative energy solutions which at first seems absurd

and artificial until further examination shows that the negative energy solutions can be interpreted as "antiparticles." So Dirac's discovery also leads to the discovery of antiparticles. A further consequence is that the vacuum is not empty but that for short times, when, according to Heisenberg, energy does not have to be conserved, there can be fluctuations in which particles and antiparticles spontaneously emerge out of the vacuum and recombine again. The "pair production" can become long lived and real when a photon of sufficient energy, a gamma ray, decomposes into an electron-positron pair. This is indeed observed experimentally. The photon energy needed twice the rest mass energies and the kinetic energies of decomposition. So the logic seems to make sense: Maxwell's equations satisfy the relativity principle by themselves, without further assumptions, but light can break up into matter and form particle-antiparticle pairs; this is also an experimental observation. But these particles must therefore necessarily also obey the relativity principle, which means they must obey the Dirac equation, when they have spin ½.

So what does this have to do with spin? It means that as the particle moves, it can for short times follow paths which are different to the normal space trajectories that we are used to. The particle can merge into an antiparticle which was spontaneously created as a quantum fluctuation and reappear as the particle component of that pair in another location. Indeed it has to do that, since for short enough time intervals, the particle still exists, but the vacuum it moves through has structure fluctuates, breaks up into matter and antimatter and reforms. Thus new pathways or "points" or space-time realizations are created and can and indeed must be passed through. For very short time intervals, the particle can visit antimatter points and form closed loops, i.e., come back again to where it was having visited antimatter points. The new "vacuum paths" look as if they are orbits and have spin angular momentum. The solutions of the Dirac equation are called fermions and have spin ½ as we have seen. The remarkable property is that even though energy corrections can disappear in the nonrelativistic limit, the spin remains, which shows that the matter-antimatter property of the vacuum still has an effect on the electron which it cannot escape. This new fabric of space time discovered by Dirac exists whether the particle has a low- or high-average velocity. In the short enough time evolution, the new space-time configurations can and are always visited. Dirac showed that the velocity of the particle is actually indeterminate; the instantaneous velocity of the particle is actually the speed of light! Then he found the reason for this strange and novel behavior by calculating the time dependence of the velocity. He discovered that the particle is undergoing an ultrafast, order of speed of light, trembling like motion with frequency $>2m_0c^2/h$, which is more than twice the rest mass frequency, and with spatial amplitude of order $\hbar/m_0c \sim 10^{-15}$ cm. We tentatively interpret these trembling motions as precisely the visits and returns into and from antiparticle space.

The solution of the Klein-Gordon equation in the simple form has apparently no spin. But it turns out that they can, and indeed, and must also have spin. They too move in a vacuum which, as Dirac showed, is not just empty space. In particular they have solutions of integer spin, the so-called bosons, for which there is no Pauli principle, but the proof which has to do with "quantum field theory" is not the subject of this book spin as an internal fabric of space time was just the beginning

elementary particle physics. Other symmetries, combined with special relativity, some intuitive others not, turn out to have similar consequences in quantum mechanics. Particles now have spin, color, charm, etc. This is the subject of the modern field of quantum chromodynamics and string theory which strive to explain the origin of mass and of gravitation in terms of the vibrational excitations of zero mass vacuum entities and then the coupling of these excitations in vacuum.

The positive energy free particle solutions to Dirac's equation are given by:

$$\Psi = \exp\left(-iEt/\hbar + i\,\vec{p}\cdot\vec{r}\right)\begin{pmatrix}1\\0\\\dfrac{cp}{E+m_0c^2}\\0\end{pmatrix} \tag{4.146}$$

$$\Psi = \exp\left(-iEt/\hbar + i\,\vec{p}\cdot\vec{r}\right)\begin{pmatrix}0\\1\\0\\-\dfrac{cp}{E+m_0c^2}\end{pmatrix} \tag{4.147}$$

$$E = \left\{m_0^2c^4 + \vec{p}\cdot\vec{p}\right\}^{1/2} \tag{4.148}$$

Note that the Pauli spin matrix form survives the nonrelativistic limit. Note also the fascinating fact that despite the time linearization, the time oscillations of the velocity, the overall wavefunction time is again only a phase! The particle once in an eigenstate stays there until disturbed. The density is again as in the nonrelativistic Schrödinger equation time independent. The connection to antiparticle space is reduced to just another angular momentum like quantum number the "spin."

4.8.1 The Electron Spin Operator

Now we know where the spin of the electron comes from, we can proceed to formulate the Pauli Dirac spin operators. The wavefunction of a fermion, i.e., a particle which obeys the Dirac equation, can be treated as a vector in a two-dimensional space so that in addition to its spatial component it also has a spin component, so that:

$$\begin{aligned}\Psi_\mu(\vec{r})|\uparrow\rangle &= \phi_\mu(\vec{r})\begin{bmatrix}1\\0\end{bmatrix}\\ \Psi_\mu(\vec{r})|\downarrow\rangle &= \phi_\mu(\vec{r})\begin{bmatrix}0\\1\end{bmatrix}\end{aligned} \tag{4.149}$$

Top one has spin up, and lower one has spin down. The operator which measures the z-component of the spin is:

$$s_z = \frac{\hbar}{2}\begin{pmatrix} 1, 0 \\ 0, -1 \end{pmatrix}. \tag{4.150}$$

The x and y components are:

$$s_x = \frac{\hbar}{2}\begin{pmatrix} 0, 1 \\ 1, 0 \end{pmatrix} \tag{4.151}$$

$$s_y = \frac{\hbar}{2}\begin{pmatrix} 0, -i \\ i, 0 \end{pmatrix} \tag{4.152}$$

The quantities in the matrix bracket are called the Pauli spin matrices.

Experimentally it was discovered that an electron also has a spin magnetic moment. The Dirac equation knows nothing about charge, so it cannot give a magnetic moment. But if we include the charge and let it move in an electromagnetic field, then one also obtains the spin magnetic moment, so that a measurement of magnetic moment corresponds to the operator:

$$m_z = \frac{q}{m_0} s_z \tag{4.153}$$

giving the values $m_z = \pm\frac{q\hbar}{2m_0}$ in MKS units. The energy of a spin in magnetic field B is the spin Zeeman coupling and is described by the term:

$$H_Z = -\underset{s}{\vec{m}}.\vec{B} \tag{4.154}$$

where:

$$m_z = \pm\frac{q\hbar}{2m_0} = \pm\frac{1}{2}\mu_B \tag{4.155}$$

The quantity μ_B is, as mentioned already, called the Bohr magneton.

4.9 The Addition of Angular Momentum

Consider an electron, in, for example, a state of angular momentum l, in an atomic orbit $n > 1$. The spin also has an effective angular momentum, so the electron now has a total angular momentum $\boldsymbol{J}$ where we can write:

$$\vec{J} = \vec{L} + \vec{s} \tag{4.156}$$

In order to see how to add angular momentum, let us consider the addition of the angular momentum of two particles with magnitudes l_1 and l_2. We write $l = l_1 + l_2$:

$$\begin{aligned}\vec{L} &= \vec{L}_1 + \vec{L}_2 \\ \vec{L} \cdot \vec{L} &= L^2 \rightarrow \hbar l(l+1)\end{aligned} \tag{4.157}$$

The allowed values of total angular momentum are given by:

$$l = |l_1 + l_2|, |l_1 + l_2 - 1|, \ldots\ldots |l_1 - l_2|. \tag{4.158}$$

So, for example, with $l_1 = 1$ and $l_2 = 2$, the allowed values are $l = 3$, $l = 2$, and $l = 1$. Each total angular momentum state has (2 l + 1) projections along the z-axis. Thus the combination $l = 3$ has the projections:

$$l_z = 3, 2, 1, 0, -1, -2, -3 \tag{4.159}$$

The same rule applies to the spin, with $l = 1$ and $s = 1/2$; the possible states of total angular momentum are $J = 3/2$, $1/2$ with projections, $Jz = 3/2, 1/2, -1/2, -3/2$ and $J = 1/2$ which gives $Jz = 1/2, -1/2$.

The addition of spins follows a similar rule. For example, two electron spins s_1 and s_2 can combine to form $S = s_1 + s_2$ with possible total spin states $S = 1$ and $S = 0$. The former is called the triplet combination and the latter the singlet. The triplet has three projections along the z-axis with $Sz = (1,0,-1)$.

4.10 The Pauli Principle Applied to Many-Electron Systems: The Slater Determinant

We have seen what the Pauli principle implies in terms of filling the energy levels of many-electron atoms and solids, but now let us consider the formal mathematical representation. The Pauli principle requires that whenever two electrons occupy the same spatial and spin eigenvalues, then the wavefunction cannot exist, i.e., it must vanish. In order to implement this rule, and in all cases where the many-electron system is not interacting, and thus wavefunctions of many electrons can be written as products of single-particle states, there is a simple and elegant representation that satisfies this condition. This representation is called the Slater determinant. The Slater determinant representation ensures that the wavefunction is antisymmetric under exchange of electron coordinates, and this in turn ensures that it vanishes when two electrons are in the same eigenstate. So let an eigenstate, for example, for the free particle system, be written as $\phi_k(r_n)\alpha(n)$ for particle r_n with spin up (α) and $\phi_k(r_n)\beta(n)$ particle r_n with spin down (β). Then a pair in k_1 and k_2 can be represented by the linear combination of eigenstates: (i) both particles have the same spin:

$$\Psi_{1,1}\left(\vec{r}_1, \vec{r}_2\right) = \frac{1}{\sqrt{2}}\left\{\phi_{k_1}(r_1)\phi_{k_2}(r_2) - \phi_{k_2}(r_1)\phi_{k_1}(r_2)\right\}\alpha(1)\alpha(2). \tag{4.160}$$

As one can see, the wavefunction has total spin = 1, i.e., is in a triplet state with $S_z = 1$ and vanishes if the spatial quantum numbers are identical. A similar state exists with both spins down corresponding to $S_z = -1$.

When the spins are opposite, we have two possible antisymmetric combinations: the state with $S = 1$, $S_z = 0$ is:

$$\Psi_{1,0}(\vec{r}_1, \vec{r}_2) = \frac{1}{\sqrt{2}}\{\phi_{k_1}(r_1)\phi_{k_2}(r_2) - \phi_{k_2}(r_1)\phi_{k_1}(r_2)[\alpha(1)\beta(2) + \alpha(1)\beta(2)] \tag{4.161}$$

and the state with $S = 0$, $S_z = 0$ is:

$$\Psi_{0,0}(\vec{r}_1, \vec{r}_2) = \frac{1}{\sqrt{2}}\{\phi_{k_1}(r_1)\phi_{k_2}(r_2) + \phi_{k_2}(r_1)\phi_{k_1}(r_2)[\alpha(1)\beta(2) - \alpha(1)\beta(2)] \tag{4.162}$$

All these four combinations are antisymmetric under exchange of coordinates. We can extend this rule for any number of electrons, and if we combine the spin (γ) and spatial quantum number k_n for particle r_n into one, and call it $q_n = (k_n, \gamma)$, we can write for an N-particle system the determinant:

$$\Psi = \frac{1}{\sqrt{N!}}\begin{vmatrix} \phi_1(\vec{q}_1), \phi_2(\vec{q}_1) \ldots\ldots\ldots\ldots \phi_N(\vec{q}_1) \\ \phi_1(\vec{q}_2) \ldots \\ \\ \phi_1(\vec{q}_N) \ldots\ldots\ldots\ldots\ldots\ldots\ldots \phi_N(\vec{q}_N) \end{vmatrix} \tag{4.163}$$

When we use Slater determinants instead of simple product of wavefunctions, we include the fact that even though the electrons are to this order strictly speaking not interacting, there is an implicit correlation in their spatial distribution which is caused by the Pauli principle. For example, same-spin electrons cannot penetrate each other; opposite spins can. This spatial correlation, in conjunction with perturbations on the system, gives rise to the so-called exchange corrections.

4.11 Summary

This completes the chapter on the principles of quantum mechanics. We have visited some of the most important and practical application of the Schrödinger equation (SE) to real physical systems. We have shown how to solve the SE for the hydrogen atom which forms the basis for the understanding the structure of all atoms. It is an amazingly powerful result even though it involves a one-electron theory. It turns out that one can use these solutions for many electrons as well, provided one allows for the screening of the nuclear charge by the presence of the other charges in some averaged way. This is the so-called mean field or self-consistent field approach, and one can and must include the antisymmetric nature of the many-electron wavefunction as expressed by the Slater determinant of Eq. (4.163). We saw how spherical symmetry gives rise to the conservation of angular momentum and how this evolves naturally out of the hydrogen atom solutions. We discussed the origin of

the electron spin and showed how that is an internal coordinate which results from the requirement of special relativity when applied to quantum mechanics. The spin quantum number brilliantly follows from the Dirac equation or what is the linearized form of the Klein-Gordon equation, a manifestation of the fact that the "empty vacuum" is only a time averaged concept. That the vacuum can, for short times, spontaneously break up into matter and antimatter and thus allow nonintuitive quantum or quantized pathways of propagation through four-dimensional space time. The "nonintuitive" pathways, symmetries, and quantum numbers are far more numerous today in modern elementary particle physics.

4.12 The Electron in a Magnetic Field

Consider what happens to the motion of an electron in a magnetic field. We will only consider the orbital motion because in the absence of spin-orbit coupling, spin and orbit motion can be treated independently. Classically the charge is subject to the electric (E) and the Lorentz force as given below, which makes the electron follow a curved path in an electric and magnetic field (B):

$$\vec{F} = -q\left(\vec{E} + \vec{v} \times \vec{B}\right) \tag{4.164}$$

where v is the velocity. Quantum mechanically we first have to derive the new Hamiltonian. We make the following observation. A classical charged particle in an electromagnetic field obeys the Hamiltonian:

$$H = \frac{1}{2m}\left(\vec{p} + q\vec{A}\right)^2 \tag{4.165}$$

where $\boldsymbol{A}$ is the vector potential. For a magnetic field $\boldsymbol{B}$, the vector potential is:

$$\vec{B} = \vec{\nabla} \times \vec{A} \tag{4.166}$$

and

$$\vec{A} = (-yB, 0, 0) \tag{4.167}$$

Giving a field in z-direction:

$$\vec{B} = (0, 0, B). \tag{4.168}$$

In quantum mechanics we generate the correct Hamiltonian simply by using the corresponding momentum operators to obtain:

$$H = \left\{ \frac{1}{2m}(p_x + qyB)^2 + \frac{p_y^2}{2m} + \frac{p_z^2}{2m} \right\} \Psi = E\Psi \tag{4.169}$$

The momenta in x- and z-direction are not coupled, so they have plane wave solutions which allow us to simplify the Schrödinger equation to the form:

$$\Psi_n(k_x, k_z) = e^{i\left(k_x x + k_y y\right)} \varphi_n(y) \tag{4.170}$$

where:

$$\frac{-\hbar^2}{2m} \frac{\partial^2 \phi_n(y)}{\partial y^2} + \left(\frac{(qB)^2}{2m} \right) (y - y_0)^2 \phi_n = \left(E - \frac{\hbar^2 k_z^2}{2m} \right) \phi_n \tag{4.171}$$

$$y_0 = -\frac{\hbar k_x}{qB} \tag{4.172}$$

We call:

$$\omega_c = \frac{eB}{m} \tag{4.173}$$

the cyclotron frequency and note that Eq. (4.171) is an equation describing a one-dimensional harmonic oscillator with the origin shifted by y_0 and for which the eigenvalues and wavefunctions are known from Sect. 4.6. The energy levels of the electron in a magnetic field are:

$$E_n = \hbar\omega_c (n + 1/2) + \frac{\hbar^2 k_z^2}{2m} \tag{4.174}$$

The magnetic levels classified under the quantum number n are called the Landau levels. The corresponding eigenstates are:

$$\Psi_n(k_x, k_z) = A_n H_n \left[\sqrt{\frac{m\omega_c}{\hbar}} (y - y_0) \right] \exp\left[-\frac{1}{2} \sqrt{\frac{m\omega_c}{\hbar}} (y - y_0)^2 + i(k_x x + k_z z) \right] \tag{4.175}$$

where the H_n as before in Sect. 4.6 are the Hermite polynomials and the A_n the normalization factors. Consider now the question of the degeneracy of each Landau level.

* *Note that if we defined the new velocity operator in the presence of a magnetic field* via *the Heisenberg equation of motion and with the Hamiltonian* Eq. (4.169), *which is the right way to define the new operator, then we would get the result*:

$$p_x = -i\hbar\frac{\partial}{\partial x} \rightarrow -ih\frac{\partial}{\partial x} + qBy \tag{4.176}$$

which shows that the correct velocity operator is now B-field dependent. This has no classical analogue. The acceleration operator can similarly be obtained by applying the Heisenberg equation of motion with the velocity operator instead of the position operator.

4.12.1 Degeneracy of the Landau Levels

We note that the energies Eq. (4.174) do not depend on the value of k_x. This implies that to every Landau level, there are many values of k_x momentum eigenstates which give the same energy. How many are there? In order to count the degeneracy, it is convenient to assume that the system is in a cubic box of size L, with periodic boundary conditions such that $\Psi(x + L, y + L, z + L) = \Psi(x, y, z)$. This condition gives rise to the momenta $k_x$which are quantized according to the rule $k_x = \frac{2\pi n_x}{L}, n_x = 0, \pm 1, \pm 2, \ldots$. Going back to Eq. (4.171), we note that the coordinate y_0 is now also limited to be in the range [0, L]. This in turn implies that the range of values:

$$y_0 = \frac{\hbar k_x}{qB_z} = \frac{\hbar n_x}{qBL} = L \tag{4.177}$$

so that the number of values of k_x with the same energy or degeneracy of the Landau level is g_L:

$$g_L = L^2\frac{qB}{\hbar} \tag{4.178}$$

In a two-dimensional x, y system with B in z-direction, we will see that for $B_z = 0$, the density of states is a constant (see Chap. 14). The crossover from the two-dimensional $B_z = 0$ spectrum to the discrete and g_L fold degenerate Landau spectrum is shown in Fig. 4.15.

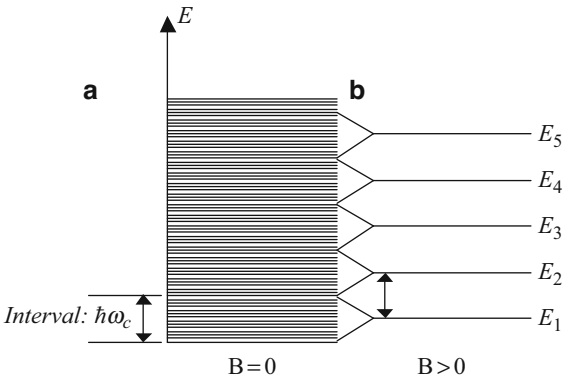


Fig. 4.15 Shows the level structure of a two-dimensional free electron system in a magnetic field. The constant density of states becomes discrete Landau levels. As the B-field increases, the degeneracy of the levels increases, and the Fermi level for a fixed electron number moves down

4.13 Discussion

In the above sections, we have considered situations where the Schrödinger equation (SE) is exactly solvable. We solved the SE in a magnetic field for spinless electrons. Let us now consider examples where an exact solution is not easily derivable analytically and where one has to resort to approximation methods. We start by formulating the very powerful Wentzel Kramer Brillouin (WKB) method.

4.14 The Wentzel Kramer Brillouin Approximation

Consider the situation where the electron moves over an arbitrary potential form $V(x)$ in the x-direction, but the motion in z and y is nearly free electron like. The wavefunction will be as:

$$\Psi(x,y,z) = A\phi(x)\exp\left[i\left(k_y y + k_z z\right)\right] \tag{4.179}$$

where A is the normalization constant and φ satisfies the one-dimensional Schrödinger equation:

$$\frac{d^2\phi(x)}{dx^2} + \frac{2m}{\hbar^2}[E - V(x)]\phi = 0 \tag{4.180}$$

$$\frac{d^2\phi(x)}{dx^2} + \frac{p^2}{\hbar^2}\phi = 0 \tag{4.181}$$

$$p = \sqrt{2m(E - V(x))} \tag{4.182}$$

where the general solution is of the form:

$$\begin{aligned}\phi(x) &= s(x)\exp\left\{\pm\frac{i}{\hbar}\int^{x} dx' p\left(x'\right)\right\} \\ &= s(x)\exp\left\{\pm\frac{i\sqrt{2m}}{\hbar}\int^{x} dx' \sqrt{E - V(x')}\right\}\end{aligned} \tag{4.183}$$

$$s(x) = Kp^{-1/2} \rightarrow K = \text{const.} \rightarrow \text{normalisation} \tag{4.184}$$

The above is called the Wentzel Kramer Brillouin (WKB) approximation and is valid when:

$$\left|\frac{h\frac{\partial p}{\partial x}}{p^2}\right| << 1 \tag{4.185}$$

$$p = \sqrt{2m[E - V(x)]} \tag{4.186}$$

In other words, when the variation of $p(x)$ or $V(x)$ is slow enough to satisfy the above condition (which is true for most situations of interest in engineering applications). In order to construct the entire solution, one considers the solution piecewise over regions of space. The most general solution in each region is of the form:

$$\phi(x) = A\frac{1}{p^{1/2}}\exp\left\{+\frac{i}{\hbar}\int^{x} p\left(x^{'}\right)dx^{'}\right\} + B\frac{1}{p^{1/2}}\exp\left\{-\frac{i}{\hbar}\int^{x} p\left(x^{'}\right)dx^{'}\right\} \tag{4.187}$$

The singular behavior at turning points when $p(E,x) = 0$ is of no serious consequence because the singularity is integrable so the wavefunction is normalizable in a finite interval.

Now one looks at a particular interval and notes that in regions where the energy E of the particle is $E > V(x)$, the function $p(x)$ is real, and the solution is a linear combination of two oscillatory waves, one going to the right and the other to the left. In regions where $E < V(x)$, $p(x)$ is complex, and the wavefunction is exponentially decaying in space in the direction against the potential barrier. The wavefunctions are multiplied by arbitrary constants A, B which have to be determined by the boundary conditions. In order to get the full solutions, one connects the different regions defined above by requiring the continuity of the wavefunction and its derivative. Consider, for example, the potential $V(x)$ in Fig. 4.16; the energy E is shown by the solid line. In region II, the energy of the particle is smaller than the potential barrier, so the wavefunction must decay exponentially as a function of $(x–x_1)$ and behave as:

$$\phi(x) = A_2\frac{1}{|p|^{1/2}}\exp\left\{+\frac{1}{\hbar}\int_{x_1}^{x_2} |p\left(x^{'}\right)|dx^{'}\right\}\} \tag{4.188}$$

In region I, between 0 and x_1, the energy E is larger than $V(x)$, so $p(x)$ is real, the phase oscillatory, and the wavefunction is of the form Eq. (4.187). In the region III, for $x > x_2$, here $E > V(x)$, and the wave has again an oscillatory structure as in Eq. (4.187):

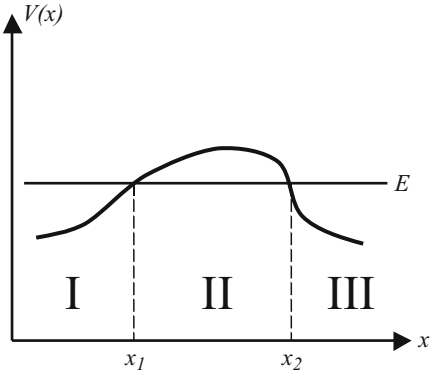


Fig. 4.16 illustrates the example treated in the text. Region II is the quantum mechanical tunneling region

$$\phi(x) = A\frac{1}{p^{1/2}}\exp\left\{+\frac{i}{\hbar}\int_{x_2}^{x} p\left(x'\right)dx'\right\} + B\frac{1}{p^{1/2}}\exp\left\{-\frac{i}{\hbar}\int_{x_2}^{x} p\left(x'\right)dx'\right\} \quad (4.189)$$

The constants A, B, A_2, can be determined by using the boundary conditions as explained above. One can in principle also determine the eigenvalues using the Wentzel Kramer Brillouin (WKB) method which is also a piecewise solution of the one-dimensional Schrödinger equation. The approximate eigenvalues can be generated by using the Bohr Sommerfeld condition which requires that the integral of $p(x)$ over the classical domain where $p(x, E)$ is real and satisfies the quantization condition:

$$\int_{p(x)>0} dx\sqrt{2m\left(E - V(x)\right)} = \left(n + \frac{1}{2}\right)\hbar \quad (4.190)$$

Though useful for finding the approximate eigenvalues of electrons confined between barriers higher than their energies, with unusual potential wells, for example, this is not the main application of WKB. The main application is when one knows the energy and one wants to know how the particle behaves in a given potential region. For example, consider the tunnel barrier as in Fig. 4.17, which is a rectangular barrier lowered by an applied field. The particle is assumed to have energy E; the question is what is the amplitude lowering when the particle has tunneled to the right to the point $p = 0$, after which it becomes an oscillatory function again. The potential in the tunnel region is $\{V(x) = V_0 - qFx\}$, so we have:

$$\phi(x) = A_3\frac{1}{|V_0 - qFx - E|^{1/2}}\exp\left\{-\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\int_{x_2}^{x}\left[V_0 - qFx' - E\right]^{1/2}dx'\right\} \quad (4.191)$$

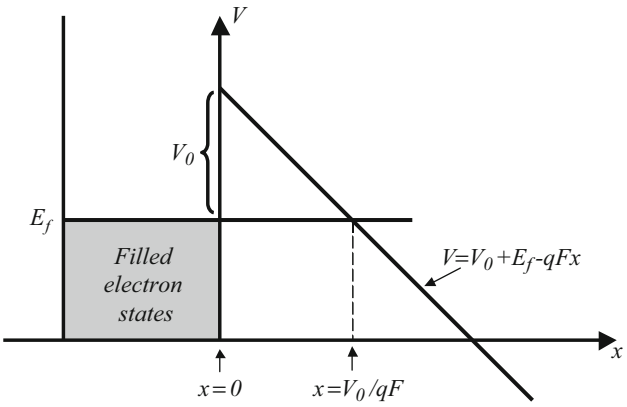


Fig. 4.17 illustrates the constant barrier in the presence of an electric field: the Fowler-Nordheim limit

The upper limit of the integral is set by the condition $E = V_0 - qFx'$, so that the decay in amplitude from the starting point of the barrier at the point x to the critical turning point is:

$$|\phi(x_c)|^2 \sim |\phi(x=0)|^2 \exp\left\{-\frac{4}{3}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\frac{(V_0-E)^{3/2}}{qF}\right\} \tag{4.192}$$

This then also gives us a measure of the transmission coefficient into the free carrier region beyond the critical point $x_c = (V_0 - E)/qF$:

$$T(E) = T_0 \exp\left\{-\frac{4}{3}\left(\frac{2m^*}{\hbar^2}\right)^{1/2}\frac{(V_0-E)^{3/2}}{qF}\right\} \tag{4.193}$$

This is called the Fowler-Nordheim tunneling structure and is encountered whenever a constant barrier is lowered by an applied electric field.

4.15 Quantum Mechanical Perturbation Theory

4.15.1 Time-Independent Perturbation

One of the most powerful results and methods of quantum mechanics is Perturbation theory. Very often one is confronted with a situation where a system is subject to an interaction or several interactions for which the complete Hamiltonian has no analytic solutions. Very frequently these interactions are also small compared to the main effect which determines the system properties. Consider, for example, a hydrogen atom in an electric field. An applied electric field of even as high as 10^7 $V \cdot cm^{-1}$ represents a tiny effect when compared to the electric field of the proton

nucleus. It is therefore very important to be able to include the effect of such new interactions, at least to some degree, and study what they do to the wavefunctions and energy levels of the system.

The procedure is as follows. We consider a starting Hamiltonian H_0 of which we assume we have the normalized eigenfunctions Φ_n and energy levels E_n. Now consider the full Hamiltonian in the presence of a perturbation V, given by $H = H_0 + V$, and we generate the solutions of this new Hamiltonian as a power series expansion, both for the energies and the wavefunctions. The expansion of energy levels and wavefunctions can be usually stopped in second order, giving us a powerful way of estimating changes in energies and wavefunctions and of course of all the relevant physical properties.

4.15.2 Nondegenerate Perturbation Theory

We first consider the situation where no two energies of the unperturbed system have the same value. Or in this particular application, we assume that the ground state is nondegenerate. In many situations of interest such as, for example, treating the effect of static electric fields on the electronic system, we can work with the time-independent Schrödinger equation and perturbation theory. The time-dependent perturbation expansion is considered in Chap. 10. We write for the new ground-state wavefunction energy the perturbation expansion:

$$\begin{aligned} \Phi_g &= \Phi_g^0 + \Phi_g^1 + \Phi_g^{(2)} \\ E_g &= E_g^0 + E_g^{(1)} + E_g^{(2)} \end{aligned} \tag{4.194}$$

The admixtures are as before linear combinations of the unperturbed eigenstates, in this case excited states, so that:

$$\Phi_g^{(n)} = \sum_{l \neq g} a_{\mathrm{lg}}^{(n)} \Phi_l^0 \tag{4.195}$$

and where the time-independent Schrödinger equation with perturbation V is given by:

$$(H_0 + V)\Phi_g = E_g \Phi_g \tag{4.196}$$

Substituting Eq. (4.194) in Eq. (4.196), and comparing coefficients of the same order, then gives to the zeroth order the obvious result:

$$H_0 \Phi_g^0 = E_g^0 \Phi_g^0 \tag{4.197}$$

First order we have:

$$H_0\Phi_g^1 + V\Phi_g^0 = E_g^0\Phi_g^1 + E_g^1\Phi_g^0 \tag{4.198}$$

In second order we have:

$$H_0\Phi_g^{(2)} + V\Phi_g^1 = E_g^0\Phi_g^{(2)} + E_g^1\Phi_g^1 + E_g^{(2)}\Phi_g^0 \tag{4.199}$$

In order to obtain the zeroth order solution, we substitute the expansion Eq. (4.195) into the first-order equation Eq. (4.198), multiply the left hand on both side by $\left(\Phi_g^0\right)*$ and integrate over all space, and use the orthogonality condition:

$$\int d\ \vec{r}\ \Phi*_n\Phi_m = \delta_{mn} \tag{4.200}$$

to find:

$$E_g^1 = \int d\ \vec{r}\ \left(\Phi_g^0\right)*V\left(\vec{r}\right)\Phi_g^0 \tag{4.201}$$

We carry on the procedure to calculate the first-order change in the wavefunction by multiplying Eq. (4.199) this time on both sides with $\left(\Phi_l^0\right)*$ and integrating while using the orthogonality again and the relation:

$$H_0\Phi_m^0 = E_m^0\Phi_m^0 \tag{4.202}$$

to find the coefficient:

$$a_{\mathrm{lg}}^{(1)} = \frac{V_{\mathrm{lg}}}{E_g - E_l} \tag{4.203}$$

and after multiplying and integrating again with $\left(\Phi_g^0\right)*$ on the second-order Eq. (4.199) and some algebra, we find the second-order energy shift:

$$E_g^{(2)} = \sum_{l\neq g}\frac{V_{gl}V_{\mathrm{lg}}}{E_g - E_l} \tag{4.204}$$

With the wavefunction given to first order by:

$$\Phi_g = \Phi_g^0 + \sum_{l\neq g}\frac{V_{\mathrm{lg}}}{E_g - E_l}\Phi_l^0 \tag{4.205}$$

whereas before the matrix element of the potential is defined by:

$$V_{ls} = \int \Phi*_s^0 V\left(\vec{r}\right)\Phi_l^0 d\ \vec{r} \tag{4.206}$$

Knowing the unperturbed wavefunctions and energy levels allows us to compute the perturbed ones. Equation (4.201), Eq. (4.204), and Eq. (4.205) are, though simple, some of the most useful results of quantum mechanics.

If, for example, we consider the particle in the one-dimensional box problem of Sect. 4.4.3, with confinement in z-direction, and we apply a perturbation which is due to an applied electric field in the z-direction, then $V = -qzE_0^z$, and we can compute the energy's shift to a good approximation with the box wavefunctions given by Eq. (4.46). If the origin is chosen as $z = L/2$ so that the box extends in the range $[-L/2 < z < L/2]$, it follows by symmetry that the first-order shift $V_{gg} = 0$, and we have only the second-order term given by:

$$E_g^{(2)} = \sum_{l \neq g} \left(qE_{0z}\right)^2 \frac{\left|z_{gl}\right|^2}{E_g - E_l} \tag{4.207}$$

If on the other hand we choose the origin to be at $z = 0$ so that the range is $[0 < z < L]$, then the expansion to second order is:

$$E_g = E_g^0 + \int_0^L dz \frac{2}{L} Sin^2\left(\frac{\pi z}{L}\right)\left(-qzE_0^z\right) + \sum_{l \neq g} \frac{\int_0^L dz \frac{2}{L} \sin\left(\frac{\pi z}{L}\right)\left(-qE_0^z z\right) \sin\left(\frac{l\pi z}{L}\right)}{E_g - E_l} \tag{4.208}$$

In summary, we have shown that to second order, the new perturbed energy levels for general perturbation V is given as:

$$E_g = E_g^0 + \int d\,\vec{r}\;\Phi_g^* V \Phi_g + \sum_{l \neq g} \frac{\left|V_{gl}\right|^2}{E_g - E_l} \tag{4.209}$$

where V_{ls} is defined by Eq. (4.206). It is interesting to note that the second-order term is for the ground-state energy, always an energy lowering term irrespective of the nature of the perturbation.

4.15.3 Degenerate-State Perturbation Theory to Second Order

When two or more energies states of the unperturbed system can have the same value, we may bypass the difficulty by using the renormalized or so-called Brillouin Wigner expansion, which to second order is the same as Eq. (4.204) except that the energy denominator contains the exact final energy and not the unperturbed ground state:

$$E_g = E_g^0 + \int d\,\vec{r}\;\Phi_g^* V \Phi_g + \sum_{l \neq g} \frac{|V_{gl}|^2}{E_g - E_l} \tag{4.210}$$

If the dominant term is due to coupling with the degenerate level $E_{g_1}^0 = E_g^0$ then to second order, the sum can be separated into the degenerate term with $l = g_1$, and the rest can stay un-renormalized to second order to give:

$$E_g = E_g^0 + V_{gg} + \frac{|V_{gg_1}|^2}{E_g - E_{g_1}^0} + \sum_{l \neq g,\, g_1} \frac{|V_{gl}|^2}{E_g^0 - E_l^0} \tag{4.211}$$

where $V_{gg} = \int d\,\vec{r}\;\Phi_g^* V \Phi_g$ and we have a simple quadratic equation to solve. Putting $E_g^0 = E_{g_1}^0$ by definition of degeneracy:

$$\left(E_g - E_g^0\right)^2 - V_t \left(E_g - E_g^0\right) - |V_{gg_1}|^2 = 0 \tag{4.212}$$

$$V_t = V_{gg} + \sum_{l \neq g,\, g_1} \frac{|V_{gl}|^2}{E_g^0 - E_l^0} \tag{4.213}$$

The quadratic has two distinct roots, and the degeneracy of the ground state is now lifted by the perturbation. The two roots are:

$$E_g = E_g^0 + \frac{1}{2}\left\{ V_t \pm \left[V_t^2 + 4V_{gg_1}^2 \right]^{1/2} \right\} \tag{4.214}$$

$$V_{gg_1} = \int d\,\vec{r}\;\Phi_g^* V(\vec{r}) \Phi_{g_1} \tag{4.215}$$

$$V_{gl} = \int d\bar{r} \Phi_g^* V(\vec{r}) \Phi_l \tag{4.216}$$

The result is very simple if we can neglect the admixture to the nondegenerate excited level or $V_{gl} = 0$; $l \neq g, g_1$.

The time-dependent perturbation method is treated in Chap. 10, in the context of optical properties, but the method presented in Chap. 10 is quite general and can be used for any time-dependent perturbation.

4.16 Final Summary

In the first part of this chapter, we introduced the principles of quantum mechanics. Then we applied the method to a number of exactly solvable problems of great physical significance: the particle in the box, the harmonic oscillator, and the hydrogen atom. We encountered angular momentum and spin. In the final parts of the chapter, we considered simple Hamiltonians which are not exactly solvable analytically and which need approximate treatments. We introduced the so-called Wentzel Kramer Brillouin (WKB) method which is a powerful method by which one can estimate the wavefunction in quasi-one-dimensional irregular potentials. We applied it to a simple but very important example with many applications: the constant potential barrier in an electric field.

In the last part of this chapter, we demonstrated how one can calculate the effect of small perturbations on quantum mechanical systems. The energy corrections were evaluated up to second order in powers of the disturbance Hamiltonian for the energy, and the corrections to the wavefunction were developed up to first order. The method was formulated for the case when the ground state is nondegenerate, and it was shown how to extend it to the case when the ground state is degenerate.

Problems

1. According to the quantum mechanics, electromagnetic radiation of frequency ν can be regarded as consisting of photons of energy $h\nu$ where $h = 6.626 \times 10^{-34}$ J·s is the Planck's constant.
   (a) What is the frequency range of visible photons (400 nm to 700 nm)? What is the energy range of visible photons (both in J and in eV)?
   (b) How many photons per second does a low power (1 mW) He-Ne laser ($336 = \lambda$ nm) emit? A cell phone that emits 0.4 W of 850 MHz radiation? A microwave oven operating at 2.45 GHz generating a microwave power of 750 W? How many photons of the latter frequency have to be absorbed to heat up a glass of water (0.2 L, heat capacity of water 4.18 $kJ \cdot kg^{-1} \cdot K^{-1}$) by 10 °C?

   At a given power of an electromagnetic wave, do you expect a classical wave description to work better for radio frequencies or X-rays? Why? At what He-Ne laser power do you expect quantum effects to become important?
2. An adapted human eye (person that has spent 30 min in the dark) can see 1 ms flashes of power 4×10^{-14} W at 510 nm with 60% reliability. Assuming that 10% of the incident power reaches the retina, how many photons at the receptors generate the signal that the test person recognizes as flash of light?
3. (a) The thermal energy scale is k_bT, where $k_b = 1.38 \times 10^{-23}$ J/K is the Boltzmann constant and T is the absolute temperature. What energy does room temperature correspond to? What would be the frequency and wavelength

of the corresponding photons? Is it reasonable that a hot body starts to glow around 1000 °C?

(b) What is the photon flux (rate of arriving photons per unit area) at 1 m distance from a 60 W light bulb, if you assume that the bulb conversion efficiency (electrical power to light bulb) is 10% and take the photon wavelength as 500 nm?

(c) A photodiode measures light power by converting incident photons into electron-hole pairs, such that the electron current is proportional to the incident light power. The quantum efficiency is defined as the probability that an incident photon generates an electron. If a typical photodiode has a responsivity of 0.5 A/W for infrared light at 850 nm, what is the quantum efficiency of the device? If the quantum efficiency is independent of frequency, what responsivity do you expect for blue light at 400 nm?

Student can find a simulation of black body radiation and related topics (Planck's law, Wien's law) at http://csep10.phys.utk.edu/guidry/java/planck/planck.html.

4. From the expression of the distribution of energy radiated by a blackbody, Eq. (4.2c) shows that the product $\lambda_M T$ is a constant, where λ_M is the wavelength of the peak of distribution at the temperature T (see Fig. 1.3).
5. Ultraviolet light of wavelength 350 nm falls on a potassium surface. The maximum energy of the photoelectrons is 1.6 eV. What is the work function of potassium? Above what wavelength will no photoemission be observed?
6. What is the deBroglie wavelength of an automobile (2000 kg) traveling at 25 miles per hour? A dust of radius 1 μm and density 200 kg·m^{-3} being jostled by air molecules at room temperature ($T = 300$ K)? An 87Rb atom that has been laser cooled to a temperature of $T = 100$ μK? An electron and a proton accelerated to 100 eV?

   Assume that the kinetic energy of the particle is given by $(3/2)k_bT$.
7. Reflection high-energy electron diffraction (RHEED) has become a commonplace technique for probing the atomic surface structures of materials. Under vacuum conditions an electron beam is made to strike the surface of the sample under test at a glancing angle ($\theta < 10°$). The beam reflects off the surface of the material and subsequently strikes a phosphorescent screen. Because of the wave-like nature of the electrons, a diffraction pattern characteristic of the first few atomic layers is observed on the screen if the surface is flat and the material is crystalline. With a distance between atomic planes of $d = 5$ Å, a glancing angle of 1°, and an operating de Broglie wavelength for the electrons of $2d\sin\theta$, compute the electron energy employed in the technique.
8. (a) Confirm, as pointed out in the text, that $<p_x> = 0$ for all energy states of a particle in a l-D box.

   (b) Verify that the normalization factor for wavefunctions describing a particle in a l-D box is $A_n = (2/a)^{1/2}$.
9. A particle with mass 6.65×10^{-27} kg is confined to an infinite square well of width L. The energy of the third level is 2.00×10^{-24} J. Calculate the value of L.

10. A particle of mass m is prepared in the ground state of an infinite-potential box of size a extending from $x = 0$ to $x = a$. Suddenly, the wall at $x = a$ is moved to $x = 2a$ within a time Δt doubling the box size. You may assume that the wavefunction is the same immediately after the change, if the change happens fast enough.
    (a) How fast is fast enough?
    (b) What is the probability that the particle is in the second ($n = 2$) state of the new well, immediately after the change? Note that the wavelength within the well, and hence the energy, for this state is the same as for the initial state in the old well. Make sure that you properly normalized wavefunctions for your calculations.
    (c) What is the probability that the particle would be found in the ground state of the sudden expansion?
    (d) What is the expectation value of the energy of the particle before and after the sudden expansion?
11. An electron is confined to a 1 micron layer of silicon. Assuming that the semiconductor can be adequately described by a one-dimensional quantum well with infinite walls, calculate the lowest possible energy within the material in units of electron volt. If the energy is interpreted as the kinetic energy of the electron, what is the corresponding electron velocity? The effective mass of electrons in silicon is 0.26 m_0, where $m_0 = 9.11 \times 10^{-31}$ kg is the free electron rest mass.
12. In examining the finite potential well solution, suppose we restrict our interest to energies where $\zeta = E/U_o < 0.01$ and permit "a" to become very large such that in Eq. (3.61), $\alpha_0 a \zeta_{max}^{1/2} >> \pi$. Present an argument which concludes that the energy states of interest will be very closely approximated by those of the infinitely deep potential well.
13. In this exercise, we will apply the material in Sect. 4.4.4 (page144) to calculate the factor of confinement of a particle in a finite well. For convenience we consider symmetric case, we will translate the x-axis so that the potential equals to 0 in the region: $-a/2 < x < a/2$.
    (a) Rewrite the Eq. (4.57) in this new coordinate system. Use the boundary condition to eliminate some trivial constants. By symmetry, we search for solutions in two families of functions: even and odd function. Show that the even solutions satisfy two equations:

$$\begin{cases} \tan\left(\dfrac{ka}{2}\right) = \dfrac{\alpha}{k} \\ k^2 + \alpha^2 = \dfrac{2mU_0}{\hbar^2} \end{cases}$$

while the odd solutions satisfy:

$$\begin{cases} -\cot\left(\dfrac{ka}{2}\right) = \dfrac{\alpha}{k} \\ k^2 + \alpha^2 = \dfrac{2mE}{\hbar^2} \end{cases}$$

How can you resolve these equations graphically?

(b) The particle is in the ground state, which is even, of energy E. Find the probability for the particle to stay in the well. This quantity is defined as the confinement factor (or coefficient of confinement).

Student can find a simulation of this problem at http://www.sgi.com/fun/java/john/wave-sim.html.

14. Consider a particle of mass m moving in the potential:

$$V(x) = -\frac{\hbar^2 a^2}{m}\frac{1}{\cosh^2(ax)}$$

(a) Show that this potential has a bound eigenstate described by the wavefunction:

$$\psi_0(x) = \frac{A}{\cosh(ax)}$$

and find the corresponding eigenenergy. Normalize ψ_0 and sketch it. This turns out to be the only bound state for this potential.

(b) Show that the wavefunction is:

$$\psi_k(x) = B\left(\frac{ik - a\tanh(ax)}{ik + a}\right)e^{ikx}$$

where $\hbar k = \sqrt{2mE}$, solves the Schrödinger equation for any positive energy E and near $\pm\infty$ the asymptotic of $\psi_k(x)$ has the plane wave form. Determine the transmission coefficient if it is defined as the square of the ratio between the amplitude of the coming wave (at $-\infty$) and that of the going out wave (at $+\infty$). What physical situation does ψ_k represents?

Student can find a simulation of this problem at http://www.kfunigraz.ac.at/imawww/thaller/visualization/vis.html.

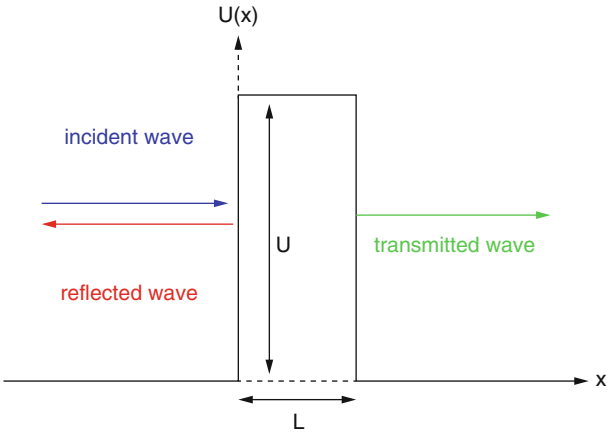
15. Using the Heisenberg equation of motion Eq. (4.30) and the Hamiltonian of a free particle in a magnetic field given by Eq. (4.171), evaluate the velocity operators vx, vy, and vz. Note how the magnetic field has modified one of the velocities. How does the presence of an electric field, if at all, modify the velocity operators?

As a consequence of relativity, the spin magnetic moment of an electron is coupled to its own orbit via the interaction:

$$H_{so} = \frac{\hbar}{4m^2c^2}\left[\vec{\nabla}\, V(\vec{r}) \times \vec{p}\right] \cdot \vec{s}$$

where $V(\vec{r})$ is the total potential seen by the electron and $\vec{s}$ is the spin operator. What is the effect of the spin-orbit interaction on the quantum mechanical definition of the velocity operator (use the Heisenberg equation of motion Eq. (4.30)). What is the dependence of the spin-orbit coupling on the orbital angular momentum L, if $V(\vec{r})$ is the Coulomb potential of the hydrogen atom?

16. A particle of energy E traveling from the left hits a barrier of height $U > E$ and thickness L. Calculate the transmission coefficient.



Student can find a simulation of this problem at http://www.kfunigraz.ac.at/imawww/thaller/visualization/vis.html http://www.sgi.com/fun/java/john/wave-sim.html.

17. The one-dimensional harmonic oscillator Eq. (4.62) is subject to an electric field F which produces an extra term qFx in the Hamiltonian. Calculate the new wavefunctions and energy levels using the zero field solutions. How does the field affect the symmetry of the charge distribution in the ground state?
18. Explain the Wentzel Kramer Brillouin (WKB) approximation. Why is it important and when would you use it? Using Eq.(4.183) verify the estimate of Eq. (4.185).
19. We have seen that in a magnetic field, the magnetic moment of an electron couples to an external magnetic field B to give the so-called Zeeman term $H_Z = -g\mu_B . B_z s_z$ where for free electrons the factor g we have introduced is called the g-value and is given without quantum field corrections by $g = -2$ and with corrections by $g = -2.0023$. In a medium the spin-orbit coupling can change the effective value of g called also the Lande's g-value. In an electron spin resonance experiment (ESR), the spins of electrons in a magnetic field can

be flipped by photon absorption. Calculate the energy of a photon needed to change the spin direction of an electron from down to up in a magnetic field of 0.3 T and a g-value of 2.35.

20. Calculate the first-order correction to the energy of an electron in electron volts eV, in the ground state of hydrogen due to the gravitational potential of the nucleus given by $V_G = -\frac{m_1 m_2 G}{r}$ where m_1 and m_2 are electron and proton masses, respectively, and G is the gravitational constant given by $G = 6.672.10^{-11}\mathrm{N{\cdot}m^2{\cdot}kg^{-2}}$.

References

Chuang L (1995) Physics of optoelectronic devices. In: Wiley series in pure and applied optics. Wiley, New York

Dirac PAM (1967) The principles of quantum mechanics, 4th edn. 'Oxford Science' Publications

Davydov AS (1965) Quantum mechanics. Pergamon, New York

Wilczek F (2006) The origin of mass. Mod Phys Lett A 21(09):701–712

Liboff RL (1998) Introductory quantum mechanics. Addison-Wesley, Reading

Further Reading

Bastard G (1988) Wave mechanics applied to semiconductor heterostructures. Halsted Press, New York

Cohen-Tannoudji C, Diu B, Laloë F (1977) Quantum mechanics. Wiley, New York

Dalven R (1990) Introduction to applied solid state physics: topics in the applications of semiconductors, superconductors, ferromagnetism, and the nonlinear optical properties of solids. Plenum Press, New York

Davydov AS (1965) Quantum mechanics. Pergamon, New York

Wilczek F (2006) The origin of mass. Mod Phys Lett A 21(09):701–712

Kittel C (1976) Introduction to solid state physics. Wiley, New York

Liboff RL (1998) Introductory quantum mechanics. Addison-Wesley, Reading

McKelvey JP (1966) Solid state and semiconductor physics. Harper and Row, New York

Pierret RF (1989) Advanced semiconductor fundamentals. Addison-Wesley, Reading

Powell JL, Crasemann B (1961) Quantum mechanics. Addison-Wesley, Reading

Ziman JM (1969) Elements of advanced quantum theory. Cambridge University Press, London

Electrons and Energy Band Structures in Crystals

5

5.1 Introduction

In Chap. 4, we introduced quantum mechanics as the proper alternative to classical mechanics to describe physical phenomena, especially when the dimensions of the systems considered approach the atomic scale. The concepts we learned will now be applied to describe the physical properties of electrons in a crystal. During this process, we will make use of the simple quantum mechanical systems which were mathematically treated in the previous chapter. This will lead us to the description of a very important concept in solid-state physics, namely, that of the "energy band structures."

5.2 Electrons in a Crystal

So far, we have discussed the energy spectrum of an electron in an atom, and more generally in a one-dimensional potential well. Modeling an electron in a solid is much more complicated because it experiences the combined electrostatic potential of all lattice ions and all other electrons. Nevertheless, the total potential acting on the electrons in a solid shares the symmetry of the lattice and thus reflects the periodicity of the lattice in the case of a crystal. This simplifies the mathematical treatment of the problem and allows us to understand how the energy spectrum, wavefunctions, and other dynamic characteristics (e.g., mass) of electrons in a solid are modified from the free particle case.

5.2.1 Bloch Theorem

The Bloch theorem provides a powerful mathematical simplification for the wavefunctions of particles evolving in a periodic potential. The solutions of the

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_5

Schrödinger equation in such a potential are not pure plane waves as they were in the case of a free particle (Eq. (4.33)) but are waves which are modulated by a function having the periodicity of the potential or lattice. Such functions are then called Bloch wavefunctions and can be expressed as:

$$\Psi\left(\vec{k}, \vec{r}\right) = \exp\left(i\,\vec{k}\,.\,\vec{r}\right).u\left(\vec{k}, \vec{r}\right) \tag{5.1}$$

where $\vec{k}$ is the wavenumber vector (in three dimensions) or wavevector of the particle, $\vec{r}$ its position, and $u\left(\vec{k}, \vec{r}\right)$ a space-dependent amplitude function which reflects the periodicity of the lattice:

$$u\left(\vec{k}, \vec{r} + \vec{R}\right) = u\left(\vec{k}, \vec{r}\right) \tag{5.2}$$

The expression in Eq. (5.1) means that the Bloch wavefunction is a plane wave, given by the exponential term in Eq. (5.1), which is modulated by a function which has the periodicity of the crystal lattice. An illustration of this is shown in Fig. 5.1 in the one-dimensional case.

Combining Eqs. (5.1) and (5.2) leads us to the form:

$$\Psi\left(\vec{k}, \vec{r} + \vec{R}\right) = \exp\left(i\,\vec{k}\,.\,\vec{R}\right)\Psi\left(\vec{k}, \vec{r}\right) \tag{5.3}$$

for any lattice vector $\vec{R}$. In a one-dimensional case, d being the period of the potential or lattice, this can be written as:

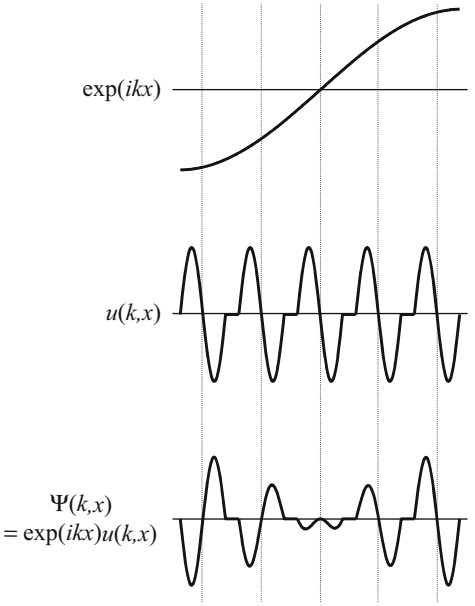


Fig. 5.1 One-dimensional illustration of a Bloch wavefunction (bottom) as a plane wave (top) modulated by a periodic function which has the period of the lattice (middle)

$$\Psi(k, x + d) = \exp(ikd)\Psi(k, x) \tag{5.4}$$

This shows that the wavefunction is the same for two values of k which differ by integral multiples of $\frac{2\pi}{d}$. We can therefore restrict the range of allowed values of k to the interval $-\frac{\pi}{d} < k \leq \frac{\pi}{d}$.

Another important limit of the Bloch theorem is for non-infinite crystals. In this case, it is common to use the periodic boundary conditions for the Bloch wavefunction, i.e., the wavefunction is the same at opposite extremities of the crystal. Assuming a linear periodic chain of N atoms (period d), the periodic boundary condition can be written as:

$$\Psi(k, x) = \Psi(k, x + Nd) = \exp(ikNd)\Psi(k, x) \tag{5.5}$$

which means that:

$$\exp(ikNd) = 1 \tag{5.6}$$

or:

$$k = \frac{2\pi n}{Nd} \tag{5.7}$$

where n is an integer. Since we restricted the range of k between $-\frac{\pi}{d}$ and $\frac{\pi}{d}$, n can only take integer values between $-\frac{N}{2}$ and $\frac{N}{2}$. There are thus only N distinct values for n and thus k.

5.2.2 One-Dimensional Kronig-Penney Model

In addition to the Bloch theorem, which simplified the wavefunction of a particle, there is a further simplification of the periodic potential, which is often used and is referred to as the Kronig-Penney model. We will continue with the one-dimensional formalism started in the previous section. In the Kronig-Penney model, the crystal is assumed to be infinite. In this model, the real crystal potential experienced by an electron is shown in Fig. 5.2a and is approximated by the one depicted in Fig. 5.2b.

The solution of the Kronig-Penney model partially utilizes the results from the finite potential well problem discussed in Sect. 4.4.4, and the same notations have therefore been used in Fig. 5.2b. The mathematical analysis will first be conducted locally, in the region $-b < x < a$, where the potential can be approximated by Eq. (4.48) except that there is a new limit for the variable x.

The wavefunction solution of the Schrödinger equation thus has two distinct components, $\Psi_1(x)$ and $\Psi_2(x)$, in different regions of space which must satisfy:

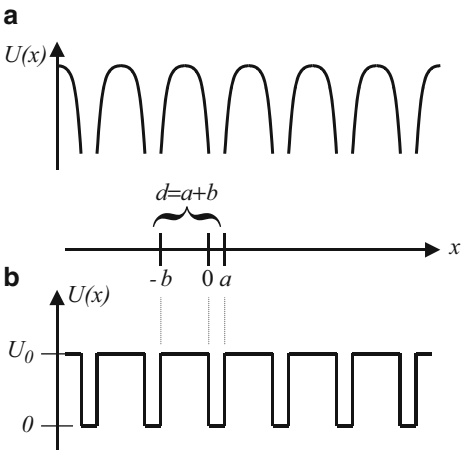


Fig. 5.2 (**a**) Real crystal potential experienced by electrons in a crystal and (**b**) simplified crystal potential used in the Kronig-Penney model

$$\begin{cases} \dfrac{d^2\Psi_1(x)}{dx^2} + \alpha^2\Psi_1(x) = 0 \quad \text{for} \quad -b < x < 0 \\ \dfrac{d^2\Psi_2(x)}{dx^2} + \beta^2\Psi_2(x) = 0 \quad \text{for} \quad 0 < x < a \end{cases} \tag{5.8}$$

by defining:

$$\begin{cases} \alpha = \begin{cases} i\alpha_-, \text{with}\, \alpha_- = \sqrt{\dfrac{2m(U_0 - E)}{\hbar^2}} & \text{when} \;\; 0 < E < U_0 \\ \alpha_+, \text{with}\, \alpha_+ = \sqrt{\dfrac{2m(E - U_0)}{\hbar^2}} & \text{when} \;\; U_0 < E \end{cases} \\ \beta = \sqrt{\dfrac{2mE}{\hbar^2}} \end{cases} \tag{5.9}$$

The general solution to Eq. (5.8) can be expressed as:

$$\begin{cases} \Psi_1(x) = A_1 \sin(\alpha x) + B_1 \cos(\alpha x) \\ \Psi_2(x) = A_2 \sin(\beta x) + B_2 \cos(\beta x) \end{cases} \tag{5.10}$$

with the understanding that $\sin(\alpha x)$ and $\cos(\alpha x)$ become $-i\sinh(\alpha_- x)$ and $\cosh(\alpha_- x)$, respectively, when the quantity $\alpha = i\alpha$ is imaginary.

The boundary conditions imply the continuity of $\Psi(x)$ and its first derivative $\frac{d\Psi(x)}{dx}$ at point $x = 0$ and include the periodicity condition of the wavefunction expressed through the Bloch theorem in Eq. (5.4) between points $x = a$ and $x = -b$:

$$\begin{cases} \Psi_1(0) = \Psi_2(0) \\ \dfrac{d\Psi_1}{dx}(0) = \dfrac{d\Psi_2}{dx}(0) \end{cases} \qquad \begin{cases} e^{ik(a+b)}\Psi_1(-b) = \Psi_2(a) \\ e^{ik(a+b)}\dfrac{d\Psi_1}{dx}(-b) = \dfrac{d\Psi_2}{dx}(a) \end{cases} \tag{5.11}$$

Utilizing Eq. (5.10), we obtain:

$$\begin{cases} B_1 = B_2 \\ \alpha A_1 = \beta A_2 \\ e^{ik(a+b)}[-A_1 \sin(\alpha b) + B_1 \cos(\alpha b)] = A_2 \sin(\beta a) + B_2 \cos(\beta a) \\ e^{ik(a+b)}[\alpha A_1 \cos(\alpha b) + \alpha B_1 \sin(\alpha b)] = \beta A_2 \cos(\beta a) - \beta B_2 \sin(\beta a) \end{cases} \tag{5.12}$$

which can be simplified by expressing A_2 and B_2 in terms of A_1 and B_1:

$$\begin{cases} A_1\left[e^{ik(a+b)} \sin(\alpha b) + \dfrac{\alpha}{\beta}\sin(\beta a)\right] + B_1\left[\cos(\beta a) - e^{ik(a+b)} \cos(\alpha b)\right] = 0 \\ A_1\left[\alpha e^{ik(a+b)} \cos(\alpha b) - \alpha \cos(\beta a)\right] + B_1\left[\beta \sin(\beta a) + \alpha e^{ik(a+b)} \sin(\alpha b)\right] = 0 \end{cases} \tag{5.13}$$

This system of two equations with two unknowns has a nonzero solution (i.e., A_1 and B_1 not both zero) if the determinant of the system is zero (for more details on the mathematics, the reader is referred to any introductory book on linear algebra). This means that the product of the first bracket in the top equation by the second bracket in the bottom equation minus the product of the second bracket in the top equation by the first bracket in the bottom equation is zero:

$$\begin{aligned} &\left[e^{ik(a+b)} \sin(\alpha b) + \frac{\alpha}{\beta}\sin(\beta a)\right]\left[\beta \sin(\beta a) + \alpha e^{ik(a+b)} \sin(\alpha b)\right] \\ &-\left[\cos(\beta a) - e^{ik(a+b)} \cos(\alpha b)\right]\left[\alpha e^{ik(a+b)} \cos(\alpha b) - \alpha \cos(\beta a)\right] = 0 \end{aligned} \tag{5.14}$$

or after simplification:

$$\cos k(a+b) = -\frac{\alpha^2 + \beta^2}{2\alpha\beta} \sin(\alpha b) \sin(\beta a) + \cos(\alpha b) \cos(\beta a) \tag{5.15}$$

Using the same constants as in Eq. (4.57), we can rewrite Eq. (5.9) as:

$$\begin{cases} \alpha = \begin{cases} i\alpha_-, \text{with}\,\alpha_- = \alpha_0\sqrt{1-\zeta} \;\; \text{when}\, 0 < E < U_0 \\ \alpha_+, \text{with}\,\alpha_+ = \alpha_0\sqrt{\zeta-1} \;\; \text{when}\, U_0 < E \end{cases} \\ \beta = \alpha_0\sqrt{\zeta} \end{cases} \tag{5.16}$$

Therefore, Eq. (5.15) can be simplified into:

$$\begin{cases} \cos k(a+b) = \dfrac{1-2\zeta}{2\sqrt{\zeta(1-\zeta)}} \sin\left(\alpha_0 a\sqrt{\zeta}\right)\sinh\left(\alpha_0 b\sqrt{1-\zeta}\right) \\ \qquad\qquad + \cos\left(\alpha_0 a\sqrt{\zeta}\right)\cosh\left(\alpha_0 b\sqrt{1-\zeta}\right) \\ \text{for } 0 < \zeta < 1 \\ \cos k(a+b) = \dfrac{1-2\zeta}{2\sqrt{\zeta(\zeta-1)}} \sin\left(\alpha_0 a\sqrt{\zeta}\right) \sin\left(\alpha_0 b\sqrt{\zeta-1}\right) \\ \qquad\qquad + \cos\left(\alpha_0 a\sqrt{\zeta}\right) \cos\left(\alpha_0 b\sqrt{\zeta-1}\right) \\ \text{for } 1 < \zeta \end{cases} \tag{5.17}$$

In these equations, the only variable in the right-hand side functions is the energy E, while the only variable in the left-hand side is the wavenumber k. Similar to the finite potential well case, a solution in ζ of Eq. (5.17) allows us to determine the values of the energy as well as the wavefunctions (after normalization).

5.2.3 Energy Bands

In the Kronig-Penney model, the crystal is assumed to be infinite. Therefore, the periodic boundary condition of the Bloch wavefunction is unnecessary, and the wavenumber k can take a continuous range of values and is real (i.e., not complex). Equation (5.17) is most easily solved graphically. The shape of the right-hand side function of Eq. (5.17), which we will call $f(\zeta)$, can be visualized in Fig. 5.3.

Because of the cosine on the LHS of Eq. (5.17), only values of $f(\zeta)$ that are between -1 and $+1$ lead to allowed (real) values for k. The areas where this occurs are shaded in Fig. 5.3. Because k is determined through a cosine function, two opposite values of k are possible for the same value for $f(\zeta)$. In these shaded areas, there is a continuous range of values for ζ (or E), corresponding to allowed energy bands. Some values of ζ, however, occur in non-shaded areas in Fig. 5.3 and are thus "forbidden," meaning that there is no possible state corresponding to these values of energy. Such regions are called regions of forbidden energy, or energy gaps. An illustration of these energy bands is given in Fig. 5.4.

Furthermore, as we can see from Fig. 5.3, for every given value of k between $-\frac{\pi}{a+b}$ and $\frac{\pi}{a+b}$, several values of ζ (thus E) are possible. An actual plot of the E–k relationship is given in Fig. 5.5 and is called the energy spectrum, the band diagram, or band structure. This type of diagram is very important in determining the

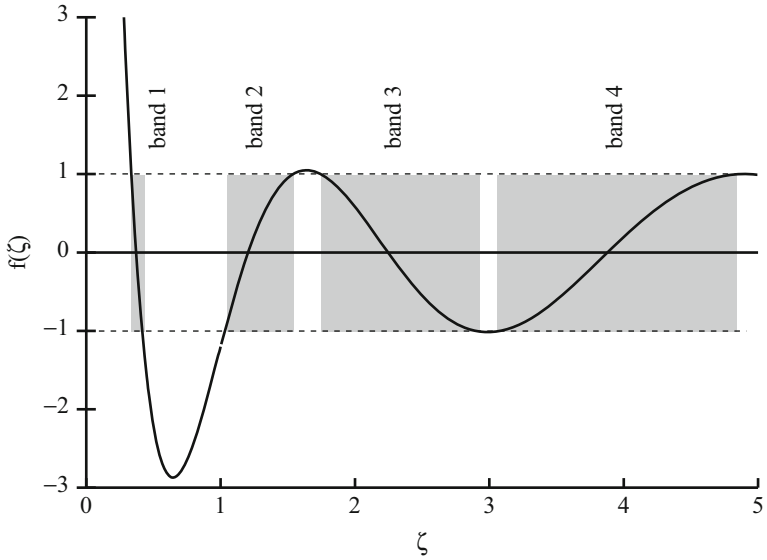


Fig. 5.3 Plot of the right-hand side of Eq. (5.17), showing the graphical determination of the E–k relationship. There exists a solution to Eq. (5.17) only when the right-hand side of the equation is between -1 and $+1$, which correspond to the shaded areas

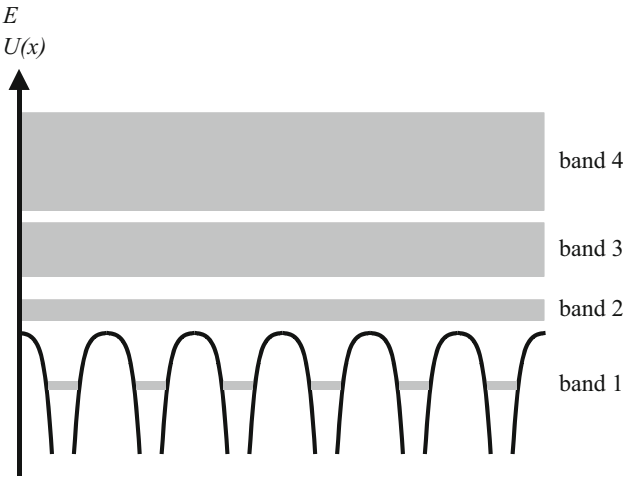


Fig. 5.4 Illustration of the concept of energy bands in the crystal

properties of an electron in a crystal. A noteworthy feature, which is true for real crystals and which can easily be seen in this diagram, is that the slope of the energy band, i.e. $\frac{dE}{dk}$, is equal to zero at the center ($k = 0$) and extremities ($k = \pm\frac{\pi}{a+b}$). This diagram, in which the value of k is restricted in the interval between $-\frac{\pi}{a+b}$ and $\frac{\pi}{a+b}$, is

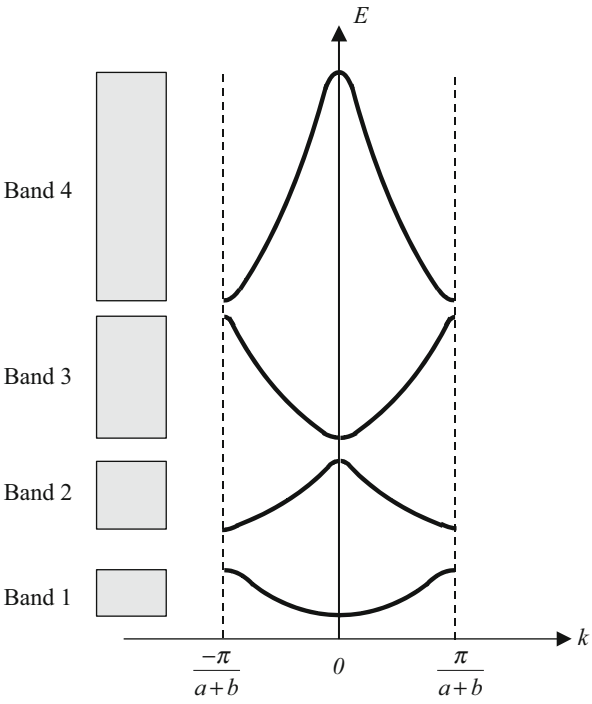


Fig. 5.5 One-dimensional E–k relationship in the reduced-zone representation in the Kronig-Penney model

often referred to as the reduced-zone representation of the energy versus k dispersion relation, as opposed to the extended-zone representation which we will now briefly discuss.

Because the energy is a periodic function of k, the reduced-zone scheme is the right way to think about the band structure of the system. All the information about the allowed energy bands is contained in the first Brillouin zone. Going outside the Brillouin zone simply repeats the same information; it does not add anything new to our knowledge. In the extended-zone representation, one can lift the previous restriction on the k-values and instead of being restricted to the values in the interval $-\frac{\pi}{a+b}$ and $\frac{\pi}{a+b}$, k is allowed to have any (larger) values. This however does not change the wavefunction because of the Bloch theorem: the k-values outside the first Brillouin zone can be reduced to ones inside the first Brillouin zone by "subtracting" a reciprocal; lattice vector $\vec{K}$. One can if one wishes unfold the band diagram into the diagram shown in Fig. 5.6, but the larger values of k can be reduced to equivalent values of k inside the first zone. Unlike for free particles, in a crystal subject to Bloch's theorem the higher values of k do not signify a higher value of momentum. Indeed, values of momentum differing from each other exactly by a reciprocal lattice vector are indistinguishable. This does not mean that $\vec{k}$ has nothing to do with momentum, it is related to the particle momentum, but it is defined and conserved only up to a reciprocal lattice vector: If one adds a reciprocal lattice vector to $\vec{k}$, the

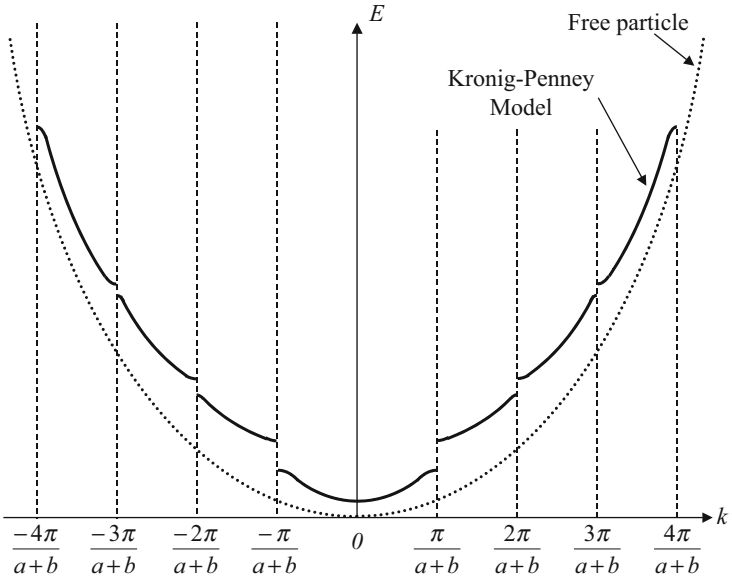


Fig. 5.6 One-dimensional E–k relationship in the extended-zone representation in the Kronig-Penney model. The parabolic relation for the free particle is shown in dotted lines for comparison. The deviation from a parabolic shape occurs mainly at the Brillouin zone boundaries

energy in the same band remains the same. The expression $\hbar k$, which corresponded to the particle momentum in the free particle case ($\langle p \rangle = \hbar k$), is now referred to as the quasi-momentum of the electron or the crystal momentum because it includes the interaction of the electron with the crystal. This explains why one can add integral multiples of $\frac{2\pi}{a+b}$ to the wavenumber without changing the band structure of the crystal, while this would be meaningless if it was a particle momentum. The reason why this quasi-momentum is not absolutely conserved in a lattice, and only conserved up to a reciprocal lattice vector, is ultimately connected to the fact that the Hamiltonian in a lattice is not translationally invariant over any arbitrary displacement as it would be in a space with no external forces, but it is only invariant when displaced by a lattice vector.

5.2.4 Nearly Free Electron Approximation

The Kronig-Penney model discussed previously is not the only method to determine the band structure in crystals, but it is the simplest and leads to a complete analytic solution. Many other methods have been developed which can be methodologically divided into two groups: one that uses the nearly free electron method and the other the tight-binding method (to be discussed below). Nevertheless, they all lead to similar results as they are merely different descriptions of the same phenomena. Here we have approximately described the band structure using the Kronig-Penney

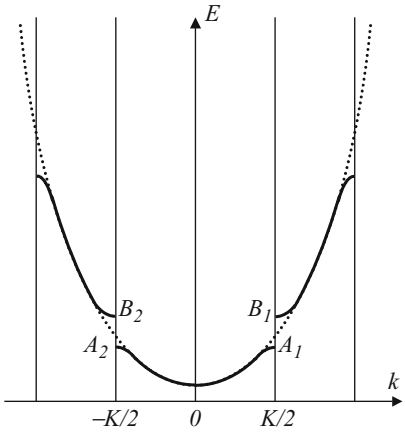


Fig. 5.7 Electron energy in a lattice (solid curve) and energy spectrum of free electrons (dashed curve). The deviation from the parabolic shape occurs at the Brillouin zone boundaries

model. In this subsection, we will briefly discuss the principle of the nearly free approximation (see Appendix A.7 for the pseudopotential approach).

This method is based on the assumption that the periodic potential introduces a small perturbation to the free electron state, i.e., a perturbation term is added to the potential energy in the Schrödinger equation, wavefunctions, and energy of the free particle to reflect this effect. Although these perturbations are small, the mathematical computation results in significant changes in the energy spectrum of a free electron. The reason is that the periodic potential scatters the electrons, and only the constructive interference of the waves survives and can propagate in the lattice as a Bloch function. The resulting band diagram in the extended-zone representation is depicted in Fig. 5.7 (solid line) and compared with that of a free electron (dashed lines).

The discontinuous curve results from the "reflections" that the electron waves with momenta of $\pm\hbar K/2$ experience at atomic lattice planes, where K is a reciprocal lattice vector (see Chap. 3 for reciprocal lattice). In the simple cubic lattice, $|K| = \frac{2\pi}{d}$ where d is the lattice constant. These locations correspond to the boundaries of the Brillouin zones defined in the previous subsection.

The energy difference between branches at points A_1 and B_1 (A_2 and B_2) is the energy gap that appears as a result of the periodic potential in the lattice. The value of the energy gap depends on the amplitude of the periodic potential. When the periodic potential reduces to be zero, the energy gaps close, and the spectrum becomes that of a free particle as shown in Fig. 4.5.

The band diagram can also be plotted in the reduced-zone representation where the energy spectrum is reduced to the smallest first Brillouin zone of range $\left[-\frac{K}{2}, +\frac{K}{2}\right]$ as shown in Fig. 5.8.

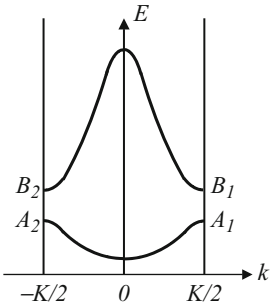


Fig. 5.8 Electron energy in the reduced-zone scheme

5.2.5 Tight-Binding Approximation

The other method commonly used to determine the band structure in a crystal, the tight-binding approximation, employs atomic wavefunctions as the basis set for the construction of the real wavefunction of an electron.

When initially isolated atoms with discrete electron energy levels are brought together and arranged in a lattice with small interatomic distances (typically ≈ 3–6 Å), the potential of each atom will be distorted due to the influence of other atoms. At the same time, the wavefunctions of electrons from different atoms will overlap, i.e., the probability of the presence of electrons from different atoms will be nonzero in the same position in space. These result in a nonzero probability for an electron to escape from one atom to the nearest neighbor. This causes a broadening of the initially discrete energy spectrum and creates energy bands of finite width instead. In other words, an electron does not live at a certain atomic energy level for an infinite time but travels from site to site which is equivalent to the movement of electrons in an energy band. Expressed mathematically, the Bloch superposition of localized orbitals gives us the tight-binding wavefunction:

$$\Psi_{\vec{k}}(\vec{r}) = \sum_{j,n} \beta_j \Phi_j(\vec{r} - \vec{R}_n) \exp(i\,\vec{k}\cdot\vec{R}_n) \tag{5.18}$$

where β_j are the admixture coefficients of the jth orbital and $\Phi_j(\vec{r} - \vec{R}_n)$ is the jth orbital itself on the atom located at $\vec{R}_n$, respectively. Substituting Eq. (5.18) into the time-independent Schrödinger equation allows us to calculate the energy bands. One does this to a good approximation by noting that the atomic problem (kinetic energy plus the potential of a given atom) is solved by the given orbital function, and the energy is known, i.e., using:

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r} - \vec{R}_i)\right\}\Phi_\alpha(\vec{r} - \vec{R}_i) = E_\alpha \Phi_\alpha(\vec{r} - \vec{R}_i)$$

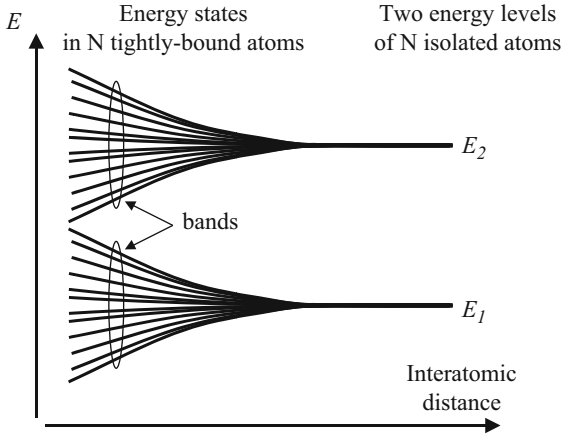


Fig. 5.9 Broadening of the atomic energy levels in a solid. When the atoms are isolated, they all have the same allowed discrete energy levels (e.g., E_1 and E_2). When the interatomic distance decreases, the atoms interact with one another and the allowed energy levels split: some increase while some others decrease

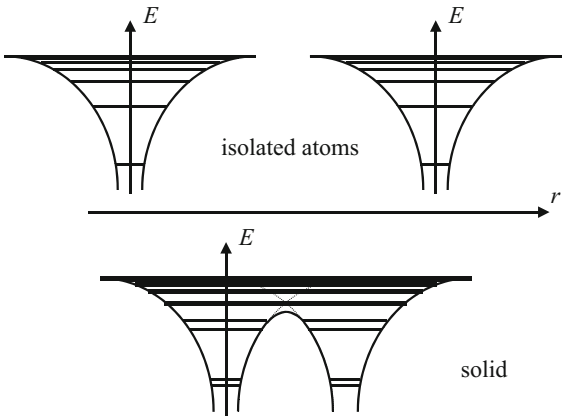


Fig. 5.10 Change in energy spectrum from single atoms to a solid. Each of the discrete energy levels in two isolated atoms split into two separate energy levels when the atoms are bound in a solid

where E_α is the energy of the atomic level and then multiplying both sides of the Schrödinger equation with a complex conjugate orbital state and then assuming the orthogonality of the orbitals centered on different sites. Normally, it is sufficient to keep only the nearest neighbor overlap terms $t_{l+1,l} = \int d\,\vec{r}\;\Phi^*\left(\vec{r}-\vec{R}_{l+1}\right)V\left(\vec{r}-\vec{R}_l\right)\Phi\left(\vec{r}-\vec{R}_l\right)$. This quantity is the so-called two-center integral, and this simplification makes the tight-binding method a good starting point for an approximate band structure calculation.

For the outer valence electrons which are usually of interest to us, the overlapping of wavefunctions is large, so the width of the energy band reaches several eV, i.e., is of the order of and even exceeds the spacing between the successive energy levels of an isolated atom. For electrons of the inner atomic shell, the level broadening is smaller, so the energy levels remain essentially sharp. The level broadening, which can be estimated to be zt where z is the number of nearest neighbors, and we take t_{ij}~t, is illustrated in Figs. 5.9 and 5.10.

Bringing atoms together and modifying their energy levels is the methodology of the "tight-binding approximation" because we start from tightly bound electrons in the atoms. This is in contrast with the previous nearly free electron approximation approach where we began with the free electron model and progressed by adding a periodic potential as a perturbation. With the tight-binding model, one arrives to a qualitatively similar band picture as that obtained from the nearly free electron model.

5.2.6 Dynamics of Electrons in a Crystal

The dynamics of electrons in a crystal can now be analyzed by considering an electron as a wavepacket. We will continue with the one-dimensional formalism of previous subsections.

Assuming that a wavepacket is centered on a frequency ω and a wavenumber k, the electron can be considered to be moving at a velocity v_g, called group velocity, which characterizes the speed of propagation of the energy that it transports. This velocity is defined by classical wave theory to be:

$$v_g = \frac{d\omega}{dk} \tag{5.19}$$

In quantum mechanics, this would correspond to the velocity of the electron. From the wave-particle duality, the frequency of the wave is related to the energy of the particle by $E = \hbar\omega$ and Eq. (5.19) thus becomes:

$$v_g = \frac{1}{\hbar}\frac{dE}{dk} \tag{5.20}$$

When an external force F acts on the wavepacket or electron so that a mechanical work is induced, it changes the energy E by the amount:

$$dE = Fdx = Fv_g dt \tag{5.21}$$

where dx is the distance over which the force is exerted during the interval of time dt. The force F can then be successively expressed as:

$$F = \frac{1}{v_g}\frac{dE}{dt} = \frac{1}{v_g}\frac{dE}{dk}\frac{dk}{dt} \tag{5.22}$$

or:

$$F = \hbar\frac{dk}{dt} = \frac{d(\hbar k)}{dt} \tag{5.23}$$

after using Eq. (5.20). On the other hand, differentiating Eq. (5.20) with respect to time leads to:

$$\frac{dv_g}{dt} = \frac{1}{\hbar}\frac{d}{dt}\left(\frac{dE}{dk}\right) = \frac{1}{\hbar}\frac{d^2E}{dk^2}\frac{dk}{dt}$$

or:

$$\frac{dv_g}{dt} = \frac{1}{\hbar^2}\frac{d^2E}{dk^2}\frac{d(\hbar k)}{dt} \tag{5.24}$$

Eliminating $\frac{d(\hbar k)}{dt}$ in Eqs. (5.23) and (5.24), we find:

$$F = \left(\frac{1}{\frac{1}{\hbar^2}\frac{d^2E}{dk^2}}\right)\frac{dv_g}{dt} \tag{5.25}$$

This expression resembles Newton's law of motion when rewritten as:

$$F = m^*\frac{dv_g}{dt} \tag{5.26}$$

where we have defined m^* as:

$$m^* = \frac{\hbar^2}{{}^{d^2E}\!/_{dk^2}} \tag{5.27}$$

m^* is called the electron effective mass and has a very significant meaning in solid-state physics. Equation (5.26) shows that, in quantum mechanics, when external forces are exerted on the electron, the classical laws of dynamics can still be used if the mass is changed in the mathematical expressions for the effective mass of the electron.

Unlike the classical definition of mass, the effective mass is not a constant but depends on the band structure of the electron. The effective mass expresses a relationship between the band structure found in previous subsections and the dynamics of an electron in a solid. This shows us how important it is to determine the band structure in the first place and that an electron in a solid is very unlike an electron in vacuum.

For example, in the case of a free electron, the energy spectrum is parabolic (Eq. (4.35)):

$$E(k) = \frac{\hbar^2k^2}{2m}$$

where m is the mass of the electron. Using Eq. (5.27), the effective mass can be found to be $m^* = m$, which means that the effective mass of a free electron is equal to its classically defined mass.

However, when the energy spectrum is not parabolic with respect to the wavenumber k anymore, as, for example, depicted in Fig. 5.7, the effective mass differs from the classical mass. We thus see that the presence of a periodic potential

results in a value of effective mass different from the classical mass. The effective mass reflects the inverse of the curvature of the energy bands in k-space (i.e., $\frac{d^2E}{dk^2}$). Where the bands have a high curvature, m^* is small, while for bands with a small curvature (i.e., almost flat bands) m^* is large.

It is also worth noticing that since $\frac{d^2E}{dk^2}$ can be negative, m^* can also be negative, although it is not interpreted so, as we will see later by considering holes (Sect. 5.3.3). A negative effective mass means that the acceleration of the electron is in the direction opposite to the external force exerted on it, as shown in Eq. (5.26). This phenomenon is possible because of the wave-particle duality: an electron has wave-like properties and can therefore be reflected from the lattice planes when its wavevector satisfies the Bragg condition. Experimentally, if the momentum given to an electron from an external force is less than the momentum in the opposite direction given from the lattice (reflection), a negative electron effective mass will be observed.

Finally, it should also be noted that experiments conducted to measure the mass of an electron only lead to an estimate of its effective mass, or at least "components" of it.

Example

Q Assuming that the energy dispersion of a band in a semiconductor can be expressed as $E = Ak^2$, where $A = 84.67$ Å^2·eV, calculate the electron effective mass in this band, in units of free electron rest mass m_0.

A We make use of the formula: $m^* = \frac{1}{\frac{1}{\hbar^2}\frac{d^2E}{dk^2}} = \frac{1}{\frac{1}{\hbar^2}\frac{d^2(Ak^2)}{dk^2}} = \frac{\hbar^2}{2A}$. In units of free electron mass, we get:

$$\begin{aligned}\frac{m^*}{m_0} &= \frac{\hbar^2}{2Am_0} \\ &= \frac{\left(1.05458 \times 10^{-34}\right)^2}{2 \times \left(84.67 \times 10^{-20} \times 1.60218 \times 10^{-19}\right)\left(0.91095 \times 10^{-30}\right)} \\ &= 0.045\end{aligned}$$

5.2.7 Fermi Energy

We have seen so far that the electron energy spectrum in a solid consists of bands. These bands correspond to the allowed electron energy states. Since there are many electrons in a solid, it is not enough to know the energy spectrum for a single electron, but the distribution of electrons in these bands must also be known to understand the physical properties of a solid. Similar to the way the electrons fill the atomic orbitals with lower energies first (Chap. 1), the electrons in a crystal fill the lower energy bands first while satisfying the Pauli exclusion principle.

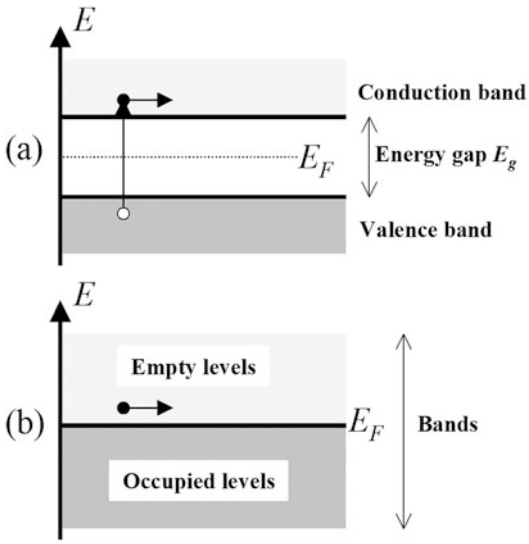


Fig. 5.11 Bands in (**a**) semiconductors and (**b**) metals. In most semiconductors, E_F is in the bandgap. In semiconductors, there is an energy region that does not contain allowed energy levels, and the Fermi energy is located in it. In metals, the Fermi energy is located inside an allowed energy band

Let us consider a solid where there are m energy levels and n electrons, at equilibrium. Usually these numbers are extremely large, and the number m of allowed energy levels (taking into account the spin degeneracy) in a solid is much larger than the number n of electrons ($m \gg n$): for instance, an iron metal with a volume of 1 cm^3 will have approximately 10^{22} atoms and 10^{24} electrons. At equilibrium, when no electron is in an excited state (e.g., at the absolute zero temperature, 0 K), the lowest n energy levels will be occupied by electrons, and the next remaining m-n energy levels remain empty.

If the highest occupied state is inside a band, the energy of this state is called the Fermi level and is denoted by E_F. That band is therefore only partially filled. This situation usually occurs for metals and is depicted in Fig. 5.11b. In the case of semiconductors, at $T = 0$ K, all bands are either full or empty. The Fermi level thus lies between the highest energy fully filled band (called valence band) and the lowest energy empty band (called conduction band), as shown in Fig. 5.11a. The energy gap between the valence band and the conduction band is called the bandgap and is denoted E_g.

The location of the Fermi level relative to the allowed energy bands is crucial in determining the electrical properties of a solid. Metals have a partially filled free electron band, since the Fermi level lies inside this band, which makes metals good electrical conductors because an applied electric field can push electrons easily into empty closely lying higher energy levels and in this way make them move in space and contribute to electrical conduction. By contrast, at 0 K, most semiconductors have completely filled or completely empty electron bands, which means that the Fermi energy lies inside a forbidden energy gap, and consequently the electric field cannot displace them from where they are in energy and therefore also not in space. Intrinsic semiconductors are poor electrical conductors at low temperatures. They

only conduct when carriers are thermally excited across the bandgap. The same can be said about insulators. Insulators differ from semiconductors in that their energy gap is much larger than k_bT, where $k_b(k_b = 1.38066 \times 10^{-23}\ \text{J}\cdot\text{K}^{-1} = 0.08625\ \text{meV}\cdot\text{K}^{-1})$ is the Boltzmann constant and T is the temperature in degrees K.

5.2.8 Electron Distribution Function

When the temperature is above the absolute zero, at thermal equilibrium, the electrons do not simply fill the lowest energy states first. We need to consider what is called the Fermi-Dirac statistics which gives the distribution of probability of an electron to have an energy E at temperature T:

$$f_e(E) = \frac{1}{\exp\left(\frac{E-E_F}{k_bT}\right) + 1} \tag{5.28}$$

where E_F is the Fermi energy and k_b is the Boltzmann constant. This distribution is called the Fermi-Dirac distribution and is plotted in Fig. 5.12 for various values of temperatures. This distribution function is obtained from statistical physics. In this description, the interaction between electrons is neglected, which is why we often talk of an electron gas.

In fact, a more general formulation of the Fermi-Dirac statistics involves a chemical potential μ instead of the Fermi energy E_F. This chemical potential depends on the temperature and any applied electrical potential. But in most cases of semiconductors, the difference between μ and E_F is very small at the temperatures usually considered.

At $T = 0$ K, the Fermi-Dirac distribution in Eq. (5.28) is equal to unity for $E < E_F$ and zero for $E > E_F$. This means that all the electrons in the crystal have their energy below E_F. At a temperature $T > 0$ K, the transition from unity to zero is less sharp. Nevertheless, for all temperatures, $f_e(E) = ½$ when $E = E_F$.

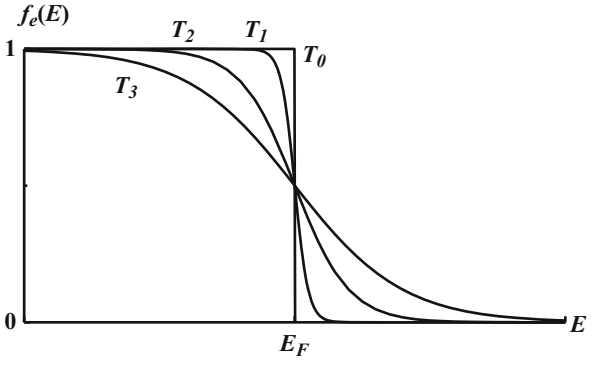


Fig. 5.12 Fermi-Dirac distribution function at different temperatures: $T_3 > T_2 > T_1, T_0 = 0$ K. At the absolute zero temperature, the probability of an electron to have an energy below the Fermi energy E_F is equal to 1, whereas its probability to have a higher energy is zero

To determine the Fermi energy, we must first introduce the concept of density of states. So far, we have somewhat indexed energy states individually, each having a certain energy. It is often more convenient to index these states according to their energy and determine the number of states which have the same energy.

5.3 Density of States (3D)

The concept of density of electronic states, or simply density of states, corresponds to the number of allowed electron energy states (taking into account spin degeneracy) per unit energy interval around an energy E. Most properties of crystals and especially semiconductors, including their optical, thermodynamic and transport properties, are determined by their density of states. In addition, one of the main motivations for considering low-dimensional quantum structures is the ability to engineer their density of states. In this section, we will present the calculation of the density of states in a bulk three-dimensional crystal, which will serve as the basis for that of low-dimensional quantum structures.

An ideal crystal has a periodic structure, which means that it has to be infinite since a surface would violate its periodicity. However, real crystals have a finite volume. We saw in Sect. 5.2 that one way to reconcile these two apparently paradoxical features in crystals was to exclude surfaces from consideration by using periodic boundary conditions (Born-von Karman). This allows us to just consider a sample of finite volume which is periodically repeated in all three orthogonal directions. A very important consequence of this was the quantization of the wavenumber k of the electron states in a crystal, as expressed through Eq. (5.7).

The analysis in Sect. 5.2 was primarily conducted in one spatial dimension (x) for the sake of simplicity. Here, it will be more appropriate to consider all three dimensions, i.e., to use $\vec{r} = (x, y, z)$.

5.3.1 Direct Calculation

Let us assume that the shape of the crystal is a rectangular parallelepiped of linear dimensions L_x, L_y, and L_z and volume $V = L_x L_y L_z$. The periodic boundary conditions, similar to Eq. (5.5), require the electron quantum states to be the same at opposite surfaces of the sample:

$$\Psi(x + L_x, y, z) = \Psi\big(x, y + L_y, z\big) = \Psi(x, y, z + L_z) = \Psi(x, y, z) \tag{5.29}$$

Using the Bloch theorem, these conditions mean that:

$$\exp\left(ik_x L_x\right) = \exp\left(ik_y L_y\right) = \exp\left(ik_z L_z\right) = 1 \tag{5.30}$$

or:

$$\begin{cases} k_x = \dfrac{2\pi}{L_1} n_x \\ k_y = \dfrac{2\pi}{L_2} n_y \\ k_z = \dfrac{2\pi}{L_3} n_z \end{cases} \tag{5.31}$$

where $n_x, n_y, n_z = 0, \quad \pm 1 \ldots$ are integers, while k_x, k_y, and k_z are the wavenumbers in the three orthogonal directions. These are in fact the coordinates of the electron wavenumber vector or wavevector $\vec{k} = (k_x, k_y, k_z)$. Therefore, the main result of the periodic boundary conditions is that the wavevector $\vec{k}$ of an electron in a crystal is not a continuous variable but is discrete. Equation (5.31) actually defines a lattice for the wavevector $\vec{k}$, and the space in which this lattice exists is in fact the k-space or reciprocal space.

The volume of the smallest unit cell in this lattice is then $\frac{(2\pi)^3}{L_x L_y L_z} = \frac{(2\pi)^3}{V}$. From Chap. 3, we know that there is exactly one lattice point in each such volume, which means that the density of allowed $\vec{k}$ is uniform and equal to $\frac{V}{(2\pi)^3}$ in k-space. Moreover, from Chap. 3, we recall that the wavevector $\vec{k}$ was used to index electron wavefunctions and therefore allowed electron states. The density of electron states per unit k-space volume is therefore equal to:

$$g(\vec{k}) = 2\frac{V}{(2\pi)^3} \tag{5.32}$$

where the extra factor of 2 arises from the spin degeneracy of electrons.

Example

Q Calculate the density of states in k-space for a cubic crystal with a side of only 1 mm. Is the density of state in k-space too low?

A The density of states in k-space is given by: $g(\vec{k}) = 2\frac{V}{(2\pi)^3} = 2 \times \frac{1\text{mm}^3}{8\pi^3} = 8.063 \times 10^{-3}$ mm^3. This number may look small, but if we compare with the volume of the first Brillouin zone, we will find that this density of states is actually very high. For example, for a face-centered cubic lattice with a lattice constant of $a = 5.65325$ Å (e.g., GaAs), the volume of its first Brillouin zone in k-space is given by: $V_k = 32\left(\frac{\pi}{a}\right)^3 = 5.492$ Å^{-3}. Therefore, the total number of possible states in this first Brillouin zone is:

$$N = V_k g(\vec{k}) = \left(5.492 \times 10^{21}\text{mm}^{-3}\right)\left(8.063 \times 10^{-3}\text{mm}^3\right) \approx 4.43 \times 10^{19}$$

The density of states $g(E)$ as defined earlier is therefore related to its counterpart in k-space, $g(\vec{k})$, by:

$$g(E)dE = g(\vec{k})d\ \vec{k} \tag{5.33}$$

where dE and $d\ \vec{k}$ are unit interval of energy and the unit volume in k-space, respectively. In order to obtain $g(E)$, one must first know the $E(\vec{k})$ relationship, which is equivalent to the E–k relationship in one dimension and which gives the number of wavevectors $\vec{k}$ associated with a given energy E. This is a critical step because the differences in the density of states of a bulk semiconductor crystal, a quantum well, a quantum wire, and a quantum dot arise from it.

For a bulk semiconductor crystal, the electron density of states is calculated near the bottom of the conduction band because this is where the electrons which give rise to the most important physical properties are located. Furthermore, we choose the origin of the energy at the bottom of this band, i.e., $E_C = 0$. Extrapolating from the results of Sect. 5.2, the shape of the $E(\vec{k})$ relationship near the bottom of the conduction band can generally be considered parabolic:

$$E(\vec{k}) = \frac{\hbar^2 k^2}{2m^*} \tag{5.34}$$

where k is the norm or length of the wavevector $\vec{k}$, and m^* is the electron effective mass as defined in Sect. 5.2.6. Using this expression, we can express successively:

$$dE = \frac{\hbar^2}{2m^*}(2k)dk \tag{5.35}$$

When considering orthogonal coordinates, the unit volume in k-space is defined given by:

$$d\ \vec{k} = dk_x dk_y dk_z \tag{5.36}$$

which is equal, when using spherical coordinates, to:

$$d\ \vec{k} = d\left(\frac{4\pi}{3}k^3\right) = 4\pi k^2 dk \tag{5.37}$$

Therefore, by replacing into Eq. (5.35), we get:

$$dE = \frac{\hbar^2}{2m^*}\left(\frac{1}{2\pi k}\right)d\ \vec{k} \tag{5.38}$$

Using Eq. (5.34) to express k in terms of E, and replacing into Eq. (5.38):

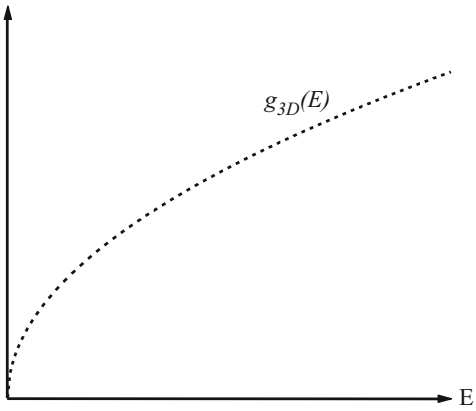


Fig. 5.13 Energy dependence of density of states for a three-dimensional semiconductor conduction band. The density of states follows a parabolic relationship

$$\begin{aligned} dE &= \frac{\hbar^2}{2m^*}\left(\frac{1}{2\pi}\sqrt{\frac{\hbar^2}{2m^*E}}\right)d\vec{k} \\ &= \frac{1}{2\pi}\left(\frac{\hbar^2}{2m^*}\right)^{3/2}\frac{1}{\sqrt{E}}d\vec{k} \end{aligned} \tag{5.39}$$

Now, by replacing into Eq. (5.33), we obtain successively:

$$g(E) = 2\pi\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E}g\left(\vec{k}\right)$$

Finally, using Eq. (5.32), we get:

$$g_{3D}(E) = \frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E} \tag{5.40}$$

where a "3D" subscript has been added to indicate that this density of states corresponding to the conduction band of a bulk three-dimensional semiconductor crystal. This density of states is shown in Fig. 5.13.

Note that, if the origin of the energies has not been chosen to be the bottom of the band (i.e., $E_c \neq 0$), then $\sqrt{E}$ would be replaced by $\sqrt{E - E_c}$.

Example

Q: Calculate the number of states from the bottom of the conduction band to 1 eV above it, for a 1 mm^3 GaAs crystal. Assume the electron effective mass is $m^* = 0.067m_0$ in GaAs.

A: The number of states from 0 to 1 eV above the bottom of the conduction band is obtained by integrating the three-dimensional density of states $g_{3D}(E)$:

$N = \int_0^{1\,eV} g_{3D}(E)dE$. Since the expression for $g_{3D}(E)$ is given by: $g_{3D}(E) = \frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E}$, we obtain:

$$
\begin{aligned}
N &= \int_0^{1\,eV} g_{3D}(E)dE = \frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\int_0^{1\,eV}\sqrt{E}dE \\
&= \frac{V}{3\pi^2}\left(\frac{2m^* \times 1eV}{\hbar^2}\right)^{3/2} \\
&= \frac{\left(10^{-3}\right)^3}{3\pi^2}\left(\frac{2\left(0.067^*0.91095 \times 10^{-30}\right) \times \left(1.60218 \times 10^{-19}\right)}{\left(1.05458 \times 10^{-34}\right)^2}\right)^{3/2} \\
&\approx 7.88 \times 10^{16}
\end{aligned}
$$

5.3.2 Other Approach

A more elegant approach, but more mathematically challenging way, to calculate the density of states is presented here. This method will prove easier when calculating the density of states of low-dimensional quantum structures. The density of states $g(E)$ as defined earlier can be conceptually written as the sum: $g(E) = 2 \times$ (number of states which have an energy $E\left(\vec{k}\right)$ equal to E) which can be mathematically expressed as:

$$
g(E) = 2\sum_{\vec{k}} \delta\left[E\left(\vec{k}\right) - E\right] \tag{5.41}
$$

where the summation is performed over all values of wavevector $\vec{k}$, since it is used to index the allowed electron states. $\delta(x)$ is a special even function, called the Dirac delta function, and is defined as:

$$
\begin{cases}
\delta(x) = 0 & \text{for } x \neq 0 \\
\int_{-\infty}^{+\infty} \delta(x)dx = 1
\end{cases} \tag{5.42}
$$

Some of the most important properties of the Dirac delta function include:

$$\begin{cases} \int_{-\infty}^{+\infty} \delta(x)Y(x)dx = Y(0) \\ \int_{-\infty}^{+\infty} \delta(x-x_0)Y(x)dx = Y(x_0) \end{cases} \tag{5.43}$$

In addition, in crystals of macroscopic sizes, the differences between nearest values of $\vec{k}$ are small, as they are proportional to $\frac{1}{L_x}, \frac{1}{L_y}$, or $\frac{1}{L_z}$. Therefore, in practice, the discrete variable $\vec{k}$ can be considered as quasi-continuous. For this reason, the summation of a function $\mathrm{Y}\left(\vec{k}\right)$ over all allowed states represented by a wavevector $\vec{k}$ in k-space can be replaced by an integration over a continuously variable $\vec{k}$ such that:

$$\sum_{\vec{k}} \mathrm{Y}\left(\vec{k}\right) \equiv \frac{V}{(2\pi)^3} \iiint_{\vec{k}} \mathrm{Y}\left(\vec{k}\right) d\,\vec{k} = \frac{V}{(2\pi)^3} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{Y}\left(k_x, k_y, k_z\right) dk_x dk_y dk_z \tag{5.44}$$

The factor $\frac{V}{(2\pi)^3}$ is the volume occupied by a reciprocal lattice point in k-space. Eq. (5.41) can therefore be rewritten into:

$$g(E) = \frac{1}{4\pi^3} \iiint_{\vec{k}} \delta\left[E\left(\vec{k}\right) - E\right] d\,\vec{k} \tag{5.45}$$

Now, we need to use the expression of $d\left[E\left(\vec{k}\right)\right]$ as a function of $d\,\vec{k}$ found in Eq. (5.39):

$$d\left[E\left(\vec{k}\right)\right] = \frac{1}{2\pi}\left(\frac{\hbar^2}{2m^*}\right)^{3/2} \frac{1}{\sqrt{E\left(\vec{k}\right)}} d\,\vec{k} \tag{5.46}$$

Equation (5.45) therefore becomes:

$$g(E) = \frac{2\pi V}{4\pi^3}\left(\frac{2m^*}{\hbar^2}\right)^{3/2} \int_0^{\infty} \delta\left[E\left(\vec{k}\right) - E\right] \sqrt{E\left(\vec{k}\right)} d\left[E\left(\vec{k}\right)\right] \tag{5.47}$$

and after the change of variable $E\left(\vec{k}\right) \rightarrow x$:

$$g(E) = \frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\int_0^\infty \delta(x-E)\sqrt{x}dx \tag{5.48}$$

Using Eq. (5.43), and because $E > 0$:

$$g(E) = \frac{V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{E} \tag{5.49}$$

which is the same expression as Eq. (5.40) for $g_{3D}(E)$.

Therefore, the knowledge of the *Fermi-Dirac distribution*, which gives us the probability of the presence of an electron with energy E, and the *density of states*, which tells how many electrons are allowed with an energy E, together permit the determination of the distribution of electrons in the energy bands. The total number of electrons in the solid, n_{total}, is therefore obtained by summing the product of the Fermi-Dirac distribution and the density of states over all values of energy:

$$n_{\text{total}} = \int_0^\infty g(E)f_e(E)dE \tag{5.50}$$

Because E_F is embedded into the function $f_e(E)$, this equation shows us how the Fermi energy can be calculated.

One important parameter for semiconductor devices is the concentration or density of electrons n in the conduction band. The following discussion provides a simplified overview of the formalism commonly used for this parameter and illustrates well the use of the Fermi-Dirac distribution. A more detailed analysis will be provided in Chap. 7 in which we will discuss the equilibrium electronic properties of semiconductors. Here, the density of electrons n, with effective mass m_c, in the conduction band is given by:

$$n = \frac{1}{V}\int_{E_C}^\infty g(E)f_e(E)dE \tag{5.51}$$

where the integration starts from E_C which is the energy at the bottom of the conduction band. In a bulk semiconductor, the density of states $g(E)$ in the conduction band is, as derived above, given by:

$$g(E) = \frac{V}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E-E_C)^{1/2} \tag{5.52}$$

Combining this expression with Eq. (5.28), the density of electrons becomes:

$$n = \frac{1}{2\pi^2}\left(\frac{2m_c}{\hbar^2}\right)^{3/2}\int_{E_C}^{\infty}(E-E_C)^{1/2}\frac{1}{\exp\left(\frac{E-E_F}{k_bT}\right)+1}dE \tag{5.53}$$

or:

$$n = N_c F_{\frac{1}{2}}\left(\frac{E_F - E_C}{k_bT}\right) \tag{5.54}$$

where:

$$N_c = 2\left(\frac{2\pi k_b T m_e}{h^2}\right)^{3/2} \tag{5.55}$$

is the effective density of states in the conduction band, and:

$$F_{\frac{1}{2}}(x) = \frac{2}{\sqrt{\pi}}\int_0^{\infty}\frac{y^{1/2}}{1+\exp(y-x)}dy \tag{5.56}$$

is the Fermi-Dirac integral. A more detailed discussion on the effective density of states and the Fermi-Dirac integral will be given in Chap. 7.

5.3.3 Electrons and Holes

We have seen that when the curvature of the *E–k* energy spectrum is positive, such as near point O in the bottom band in Fig. 5.8, the electron effective mass is positive.

However, when the curvature is negative, such as near point A_1 in this same band, the effective mass of the electron as calculated in Sect. 5.2.6 would be negative. In this case, it is more convenient to introduce the concept of holes. A hole can be viewed as an allowed energy state that is non-occupied by an electron in an almost filled band. Figures 5.14a, b are equivalent descriptions of the same physical phenomenon. In Fig. 5.14a, we are showing the energy states occupied by electrons. In Fig. 5.14b, we are showing the energy states in the valence band which are occupied by holes, i.e., vacated by electrons.

Electrons can move in such a band only through an electron filling this non-occupied state and thus leaving a new non-occupied state behind. By doing so, it is as if the vacated space or hole had also moved, but in the *opposite direction*, which means that the effective mass of the hole is therefore opposite that of the electron that would be at that same position, in other words, the effective mass of the hole is positive near point A_1 in Fig. 5.8 and is computed as:

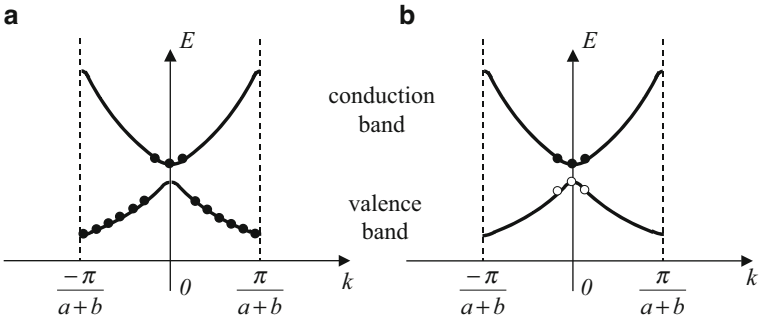


Fig. 5.14 Electron energy states in the reduced-zone scheme. In (**a**), the solid circles show the states occupied by electrons. In (**b**), the closed circles show the states in the conduction band which are occupied by electrons, and the open circles the states in the valence band occupied by holes

$$m^* = -\frac{\hbar^2}{d^2E/dk^2} \tag{5.57}$$

A hole can be viewed as a positively charged particle (energy state vacated by an electron). Holes participate in the electrical charge transfer (electrical current) and energy transfer (thermal conductivity).

Let us consider the concept of holes in more details. The probability of the state k to be occupied by an electron is $f_e(k)$. The probability of the state not to be occupied is the probability to find hole in the state k and can be written as:

$$f_h(k) = 1 - f_e(k) \tag{5.58}$$

The electrical current from the electrons in the band is:

$$j = -2q \sum_k f_e(k) v_k \tag{5.59}$$

where ν_k is the electron velocity at state k, q is the electron charge ($q > 0$) and the summation is performed over all states with wavenumber k in the first Brillouin zone. This can be rewritten as:

$$\begin{aligned} j &= -2q \sum_k f_e(k) v_k = -2q \sum_k [1 - f_h(k)] v_k \\ &= -2q \sum_k v_k + 2q \sum_k f_h(k) v_k \end{aligned} \tag{5.60}$$

We can now use the fact that the electron energy spectrum is always symmetrical, i.e., $E(k) = E(-k)$; hence $v_k = -v_{-k}$ from Eq. (5.20), and the sum of velocities over the entire first Brillouin zone is zero. The first sum in Eq. (5.60) is thus equal to zero and we obtain:

$$j = +2q \sum_k f_{\mathrm{h}}(k) v_{\mathrm{k}} \tag{5.61}$$

Therefore, the electrical current in a band incompletely filled with electrons moving at speed v_{k} is equivalent to the current of positively charged holes moving at speed v_{k}. We thus see that in a band incompletely filled with electrons, the electrical current can be represented by flow of positively charged particles-holes.

5.4 Band Structures in Real Semiconductors

In three-dimensional crystals with three-dimensional reciprocal lattices, the use of a reduced-zone representation is no longer merely a convenience. It is essential; otherwise, the representation of the electronic states becomes too complex. How then can we display the band structure information from a three-dimensional crystal, which needs of course four dimensions (E, k_x, k_y, and k_z) to describe it? The answer is to make representations of certain important symmetry directions in the three-dimensional Brillouin zone as one-dimensional E versus k plots. Only by doing so can we get all the important information onto a two-dimensional page. Therefore, when looking at an E–k diagram, one is looking at different sections cut out of the k-space. In addition, to simplify the diagram, we consider that k varies continuously. Indeed, the difference between two values of k is $\Delta k = \frac{2\pi}{Na}$, where the lattice parameter a is around several angstroms and the order of magnitude of N is 10^8. And the length of the side of the Brillouin zone is $\frac{2\pi}{a} \approx 6.28 * 10^{10}\ \mathrm{m}^{-1} \gg \frac{2\pi}{Na} \approx 6.28 * 10^{10}\ \mathrm{m}^{-1}$. As a result, at the scale of the reciprocal lattice, the wavenumber can be considered to vary continuously.

5.4.1 First Brillouin Zone of an fcc Lattice

The first Brillouin zone of an fcc lattice is shown in Fig. 5.15. Certain symmetry points of the Brillouin zone are marked. Roman letters are mostly used for symmetry points and Greek letters for symmetry directions, specifically the Γ, X, W, K, and L points and the directions Δ, Λ, and Σ. The following is a summary of the standard symbols and their locations in k-space, with a the side of the conventional cubic unit cell:

$$\Gamma \quad \frac{2\pi}{a}(0,0,0)$$

$$\mathrm{X} \quad \frac{2\pi}{a}(0,0,1)$$

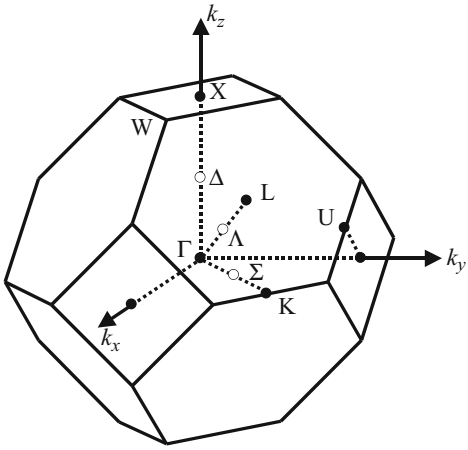


Fig. 5.15 First Brillouin zone of an fcc lattice

$$\text{W} \quad \frac{2\pi}{a}(½, 0, 1)$$

$$\text{K} \quad \frac{2\pi}{a}(¾, ¾, 0)$$

Note that there may be several equivalent positions for each of these points. For example, there are six equivalent X symmetry points, located at coordinates $\frac{2\pi}{a}$ (0,0,±1), $\frac{2\pi}{a}$(0, ±1,0), and $\frac{2\pi}{a}$(±1,0,0).

Using Miller indices, the symmetry directions can be denoted as:

$$\Delta: \quad \Gamma \rightarrow \text{X (parallel to } < 100 >)$$

$$\Lambda: \quad \Gamma \rightarrow \text{L (parallel to } < 111 >)$$

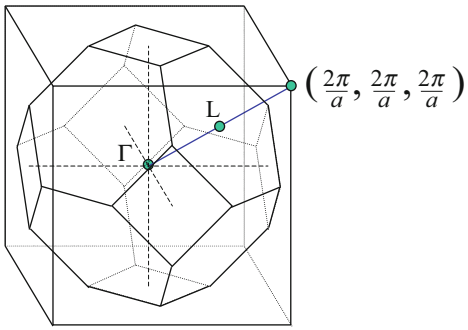
$$\Sigma: \quad \Gamma \rightarrow \text{K (parallel to } < 110 >).$$

These notations come from the crystal group theory where they are used to label the symmetry operation groups at those particular high-symmetry points and directions. For example, Γ is the symmetry group at the zone center ($\vec{k} = (0, 0, 0)$) and is isomorphic to the lattice point group.

Example

Q: Determine the coordinates of the L point in the first Brillouin zone of a face-centered cubic lattice.

A: The first Brillouin zone of a face-centered cubic lattice with side a is body-centered cubic with a side equal to $\frac{4\pi}{a}$ in the k_x, k_y, and k_z directions, as shown in the figure below. Let us take the Γ point at the center of the first Brillouin zone. The L point is exactly at the bisection point of Γ and the lattice point at ($\frac{2\pi}{a}$,$\frac{2\pi}{a}$,$\frac{2\pi}{a}$). Its coordinates are thus: ($\frac{\pi}{a}$,$\frac{\pi}{a}$,$\frac{\pi}{a}$).



5.4.2 First Brillouin Zone of a bcc Lattice

Similarly, the first Brillouin zone of a bcc lattice can be described in terms of its principal symmetry directions as it is shown in Fig. 5.16.

The symmetry points are conventionally represented as Γ, H, P, and N, and the symmetry directions as Δ, Λ, D, Σ, and G. The various symmetry points are:

$$\Gamma \quad \frac{2\pi}{a}(0,0,0)$$

$$\mathrm{H} \quad \frac{2\pi}{a}(0,0,1)$$

$$\mathrm{P} \quad \frac{2\pi}{a}(½,½,½)$$

$$\mathrm{N} \quad \frac{2\pi}{a}(½,½,0).$$

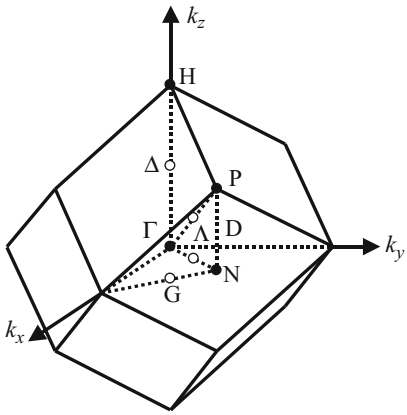


Fig. 5.16 First Brillouin zone of a bcc lattice

Using Miller indices for the directions:

$$\Delta: \quad \Gamma \rightarrow \mathrm{H}\ (\text{parallel to } < 100 >)$$

$$\Lambda: \quad \Gamma \rightarrow \mathrm{P}\ (\text{parallel to } < 111 >)$$

$$\mathrm{D}: \quad \mathrm{N} \rightarrow \mathrm{P}\ (\text{parallel to } < 100 >)$$

$$\Sigma: \quad \Gamma \rightarrow \mathrm{N}\ (\text{parallel to } < 110 >)$$

$$\mathrm{G}: \quad \mathrm{N} \rightarrow \mathrm{H}\ (\text{parallel to } < 1\bar{1}0 >).$$

5.4.3 First Brillouin Zones of a Few Semiconductors

As discussed in Chap. 3, many semiconductors have the diamond or zinc blende lattice structures. In these cases, the extrema in the E–k relations occur at the zone center or lie, for example, along the high-symmetry Δ (or <100>) and Λ (or <111>) directions. The important physical properties involving electrons in a crystal can thus be derived from plots of the allowed energy E versus the magnitude of k along these high-symmetry directions.

Figure 5.17 depicts the E–k diagrams characterizing the band structures in Ge (Fig. 5.17a), Si (Fig. 5.17b), and GaAs (Fig. 5.17c). The lines shown here represent bands in the semiconductor. The three lower sets of lines correspond to the valence band, while the upper bands correspond to the conduction bands. Note that the energy scale in these diagrams is referenced to the energy at the top of the valence band, E_V is the maximum valence band energy, E_C the minimum conduction-band energy, and$E_g = E_C$–E_V the bandgap. This is only a conventional choice, and the origin of energy can be chosen elsewhere.

The plots in Fig. 5.17 are two-direction composite diagrams. The <111> direction is toward point L, and the <100> direction is toward point X. Because of crystal symmetry, the $-\vec{k}$ portions of the diagrams are just the mirror images of the corresponding $+\vec{k}$ portions. It is therefore standard practice to delete the negative portions of the diagrams. The left-hand portions($\Gamma \rightarrow$ L) of the diagrams are shorter than the right-hand portions ($\Gamma \rightarrow$ X) as expected from the geometry of Brillouin zone.

Valence Band

In all cases, the valence band maximum occurs at the zone center, at $k = 0$. The valence band in each of the materials is actually composed of three subbands. Two of the bands are degenerate (have the same energy) at $k = 0$, while the third band is split from the other two. In Si, the upper two bands are almost indistinguishable in Fig. 5.17b and the maximum of the third band is only 0.044 eV below E_V at $k = 0$.

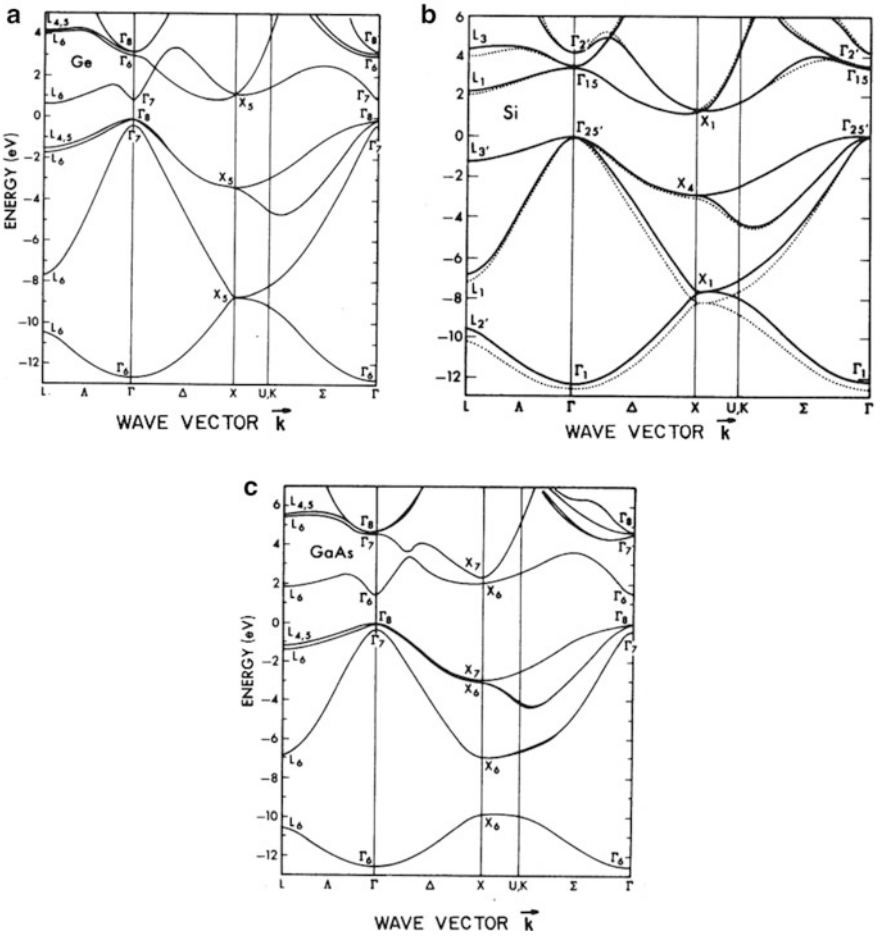


Fig. 5.17 *E–k* diagram of a few semiconductor crystals: (**a**) Ge, (**b**) Si, and (**c**) GaAs. The structures of the conduction and valence bands are plotted. The origin of the energy is chosen to be at the top of the valence band. (Reprinted figure with permission from Chelikowsky and Cohen (1976). Copyright 1976 by the American Physical Society)

The degenerate band with the smaller curvature about $k = 0$ is called the heavy-hole band, and the other with larger curvature is called the light-hole band. The band maximizing at a slightly reduced energy is called the spin-orbit split-off band (see the Kane effective mass method in Sect. 5.6).

Conduction Band

There are a number of subbands in each of the conduction bands shown in Fig. 5.17. These subbands exhibit several local minima at various positions in the Brillouin zone. However – and this is very significant – the position of the conduction band

absolute minimum in *k*-space, which is the lowest minimum among all these subbands and which is where the electrons tend to accumulate, varies from material to material.

In Ge the conduction band (absolute) minimum occurs right at point *L*, the zone boundary along the Λ or <111> direction in Fig. 5.17a. Actually, there are *eight equivalent conduction band minima* since there are eight equivalent <111> directions. However, each minimum is equally shared with the neighboring zone, and there is therefore only a *fourfold degeneracy* or a *multiplicity of four*. The other local minima in the conduction band occurring at higher energies are less populated and are therefore less important.

The Si conduction band absolute minimum occurs at $k \approx 0.8(2\pi/a)$ from the zone center along the Δ or <100> direction. The sixfold symmetry of the <100> directions gives rise to *six equivalent conduction band minima* within the Brillouin zone. The other local minima in the Si conduction band occur at considerably higher energies and are typically not important as they would only have a negligible electron population unless some very strong force could activate carriers to these higher extrema or if the temperature is much higher.

Among the materials considered in Fig. 5.17, GaAs is unique in that the conduction band minimum occurs at the zone center directly over the valence band maximum. Moreover, the L-valley minimum at the zone boundary along the <111> directions lies only 0.29 eV above the absolute conduction band minimum at Γ. Even in thermal equilibrium at room temperature, the L-valley contains a non-negligible electron population. The transfer of electrons from the Γ-valley to the L-valley can, for example, happen at high electric fields when electrons are heated up to high velocity. The transfer keeps the high energy but gives them a high effective mass which slows them down in space. When they slow down, they force the new electrons coming in to slow down too, until they, the transferred valley charge has exited. This results in a self-oscillating current state and is an essential feature for some device operations such as in charge-transferred electron devices (e.g., Gunn diodes, etc.).

Having discussed the properties of the conduction and valence bands separately, we must point out that the relative positions of the band extreme points in *k*-space are in itself an important material property. When the conduction band minimum and the valence band maximum occur at the same value of *k*, the material is said to be direct-gap type. Conversely, when the conduction band minimum and the valence band maximum occur at different values of *k*, the material is called indirect-gap type.

Of the three semiconductors considered, GaAs is an example of a direct-gap material, while Ge and Si are indirect-gap materials. The direct or indirect nature of a semiconductor has a very significant effect on the properties exhibited by the material, particularly its optical properties. The direct nature of GaAs, for example, makes it ideally suited for use in semiconductor lasers and infrared light-emitting diodes.

5.5 Two-Dimensional Semiconductors and Transition Metal Dichalcogenides "TMDC"

5.5.1 Examples: Graphene (G) and TMDC

One of the great discoveries of recent times is the exfoliation of the material which has been named graphene (G) [Geim, Castro, Avouris]. It was fabricated initially by exfoliation from graphite and consists ideally of a single carbon sheet. The good news is that this monolayer of carbon is strong enough to survive experimental manipulation and temperature, even in suspension, and can therefore be used in making devices. Graphene turned out to be so interesting that the discoverers Novoselov and Geim were awarded the Noble prize. The literature on graphene is now huge, and we will here only focus on a few noteworthy aspects which have enriched semiconductor science. The interested reader is strongly urged to consult the vast literature.

The structure of G is hexagonal two-dimensional carbon and shown in Fig. 5.18.

5.5.2 Graphene Band Structure: Nearest Neighbor Tight Binding

The simplest and most popular way of deriving the graphene band structure is to use the tight-binding method described in our *Appendix 2 in Chap. 2*. Using A and B to denote the two types of atoms which form the hexagonal lattice (see Fig. 5.18), we can assign a valence orbital to every carbon atom and allow these orbitals to couple to generate the graphene energy bands. Consider the t-b (tight-binding) Hamiltonian for this lattice and drop the Coulomb interaction between the electrons, steps which can be justified later. Using second quantization, and the creation $c_{i\sigma}^{+}$ and annihilation $c_{i\sigma}$ operators for electrons at atomic orbitals at a given site "i" [Da Sarma et al, Castro et al] [see Chap. 16 electron phonon interaction for second quantization]:

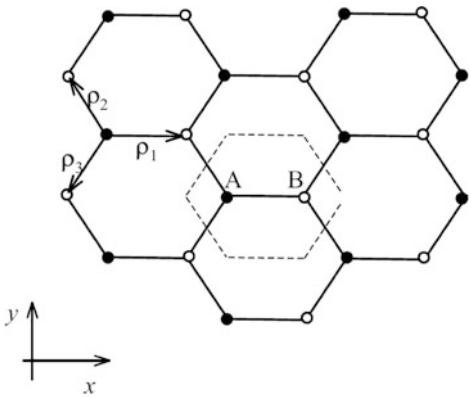


Fig. 5.18 Hexagonal lattice of a 2D material

$$H_e = \sum_{i,\sigma} \varepsilon_{i\sigma} c_{i\sigma}^{+} c_{i\sigma} + \sum_{i \neq j,\sigma} t_{ij} c_{i\sigma}^{+} c_{j\sigma} \tag{5.62}$$

$\varepsilon_{i\sigma}$ are the atomic orbital energies with spin index σ, and "t" is the tight-binding coupling matrix element linking two neighboring orbital orbitals.

Now let us set up the Heisenberg equation of motion [see FSSE Chap. 4] for the amplitude of the wavefunction at atomic sites or corresponding operator with atom of type "a," for example, E = energy:

$$(E - \varepsilon_{i\sigma}) c_{i\sigma}^{a} = \sum_{j} t_{ij} c_{j\sigma}^{b} \tag{5.63}$$

The sum j goes over the n.n and we note that the neighboring atoms a, b are *not equivalent by translational symmetry*, though apparently physically completely equivalent, so that the Bloch periodicity argument cannot be used straight away. In order to recover an *equivalent atom* and solve the equations by symmetry, we have to go one step further and set up a similar relation for the b-sites as we did for the a-sites with Eq. 5.63, thus we have:

$$(E - \varepsilon_{i\sigma}) c_{i\sigma}^{b} = \sum_{j} t_{ij} c_{j\sigma}^{a} \tag{5.64}$$

Now, we substitute Eq. (5.64) in Eq. (5.63) and relate two equivalent **a** or **b** atoms at distance **R** by the Bloch's phase factor exp(**ik.R**) in the usual way we can solve the problem. The unusual linear dispersion at the points in **k** space **K and K′** as shown in Fig. 5.19 now called Dirac points, where the gap is zero, is due to the fact that the two sets of lattice points "a and b" are completely equivalent apart from the fact that they are mirror symmetrical not translationally symmetrical. This topological restriction, analogous to the restriction of having to move below light speed at all times, then gives rise to space splitting and a new pseudo spin-like quantum number. One can see that this notion can be generalized to an infinity of topologies, which could have this, and indeed far more complex chiral symmetries. Energies can be degenerate, but one level can be "hole like" and the other "particle like."

Another way to derive the band structure is to use the real spatial wavefunction ψ and then the Wallace expectation value and optimization process [P R Wallace Phys Rev. 71, 622, (1947)]:

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathrm{A}} \exp(i\mathbf{k} \cdot \mathbf{R}_{\mathrm{A}}) X(\mathbf{r} - \mathbf{R}_{\mathrm{A}}) + \lambda \sum_{\mathrm{B}} \exp(i\mathbf{k} \cdot \mathbf{R}_{\mathrm{B}}) X(\mathbf{r} - \mathbf{R}_{\mathrm{B}}), \tag{5.65}$$

r is the particle coordinate, $\mathbf{R_A}$ and $\mathbf{R_B}$ are position of the two types of atoms (see Fig. 4.7), and **k** is the Bloch wavevector:

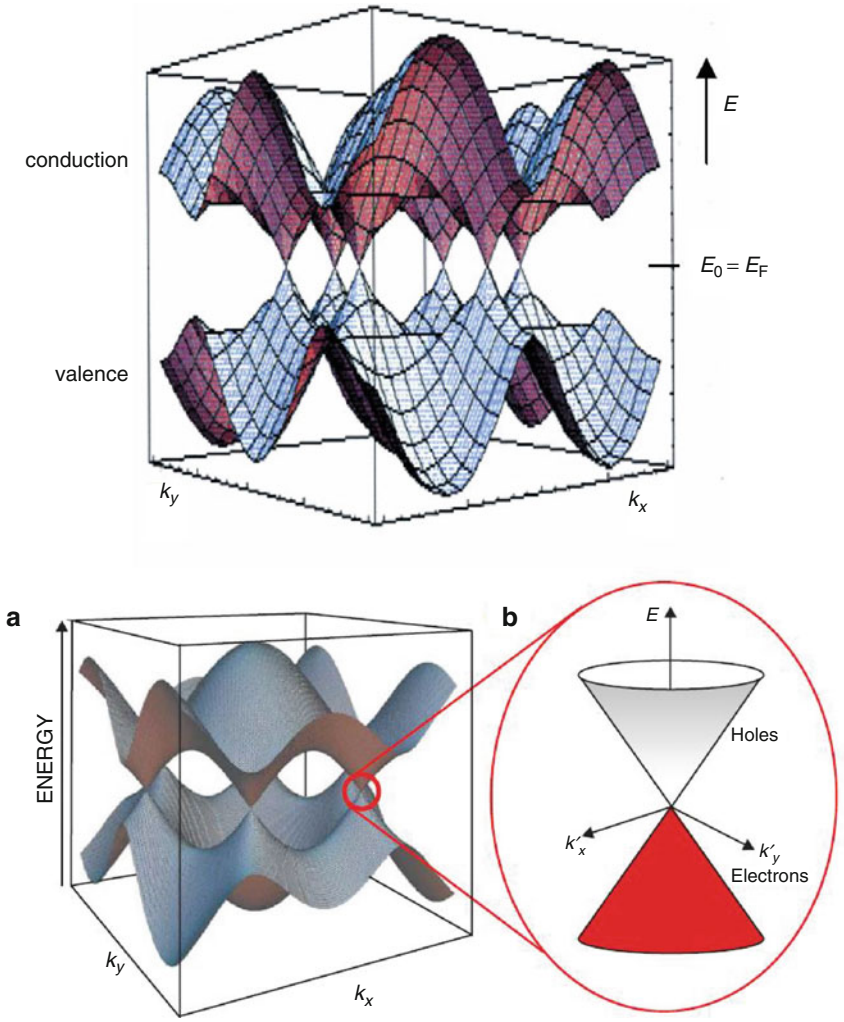


Fig. 5.19 *From Phaedon Avouris.* "Graphene electronic and photonic properties and devices" Nanoletters vol 10, p. 4285, (2010)

$$t_i = \int X^*(\mathbf{r} - \mathbf{R}_A) H X(\mathbf{r} - \mathbf{R}_{B,i}) d\mathbf{r},$$
$$E_{\mathbf{k}} = E_0 \pm \left| \sum_i t_i \exp(-i\mathbf{k} \cdot \boldsymbol{\rho}_i) \right|, \tag{5.66}$$

$\lambda = 1$ or -1

Band structure of graphene plotted in 3D to exhibit the zero gap or Dirac points K,K′.

Analytic tight-binding band structure is:

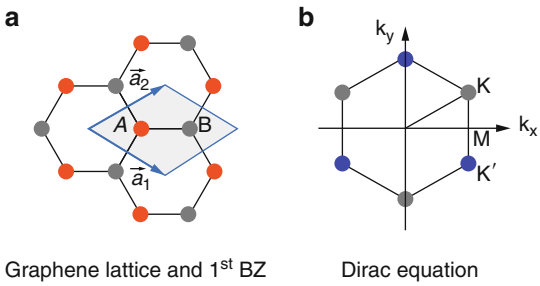


Fig. 5.20 Near the K K's points, shown above, now called Dirac points, the dispersion (energy momentum relation) is linear implying a zero effective mass

$$E(k) = \pm\gamma_0\sqrt{1 + 4\cos\left(\frac{3}{2}k_x a\right)\cos\left(\frac{\sqrt{3}}{2}k_y a\right) + 4\cos^2\left(\frac{\sqrt{3}}{2}k_y a\right)}, \tag{5.67}$$

where γ_0 is the banding energy t (atom to atom overlap) (Fig. 5.20).

The honeycomb structure can be thought of as a triangular lattice with a basis of two atoms per unit cell with 2D lattice vectors $\mathbf{A_0} = (a/2)(3, \sqrt{3})$ and $\mathbf{B_0} = (a/2)(3, -\sqrt{3})$, where $a = 0.142$ nm is the carbon-carbon distance, and $\mathbf{K} = (2\pi/(3a), 2\pi/(3\sqrt{3}\,a))$ and $\mathbf{K}' = (2\pi/(3a), -2\pi/(3\sqrt{3})\,a)$ as the inequivalent corners of the BZ and are called "Dirac points." The Dirac points play a role similar to the role of Γ points in direct bandgap semiconductors.

Relative to the Dirac point, the dispersion is:

$$E_{\pm}(q) = \pm\hbar v_F q + O(q/k)^2 \tag{5.68}$$

The dispersion depends on the Fermi velocity v_F. In tight binding v_F can be expressed in terms of the nearest neighbor hopping integral t so that:

$$\begin{aligned} &\hbar v_F = \frac{3ta}{2} \\ &a = 0.14 \text{ nm}, \quad t = 2.5 \text{ eV}, \quad v_F = 10^8 \text{ cm/s} \end{aligned} \tag{5.69}$$

The linear dispersion is like the dispersion of light or photons with:

$$E = c\hbar q \tag{5.70}$$

where c is the velocity of light. But there are here two sublattices A, B in the structure of G which allows us to write the Hamiltonian on the two sides of the bandgap as a relativistic Dirac-like Hamiltonian:

$$H = v_F \sigma \bullet \hbar q \tag{5.71}$$

where σ is a spinor-like wavefunction, v_F is the Fermi velocity of G, and $\mathbf{q}$ is the wavevector of the electron. Linear dispersion can be thought of as zero effective mass. The spinor nature of the wavefunction is not a consequence of electron spin as in the Dirac equation, but rather from the fact that there are two atoms per unit cell A,

B, and the electron can be thought of as jumping between the components A, B which is then analogous to having a pseudo spin coordinate (in the Dirac equation, the electron can be thought of as hopping into its antiparticle and back again on the time scale short enough that it is allowed by Heisenberg uncertainty principle). Whereas in the latter case, the energy to be overcome is the energy to create a particle-hole pair ($2mc^2$) in empty space, the great new effect here is that the bandgap is zero, so that the particle and hole can be created with zero energy cost. This exciting property implies that one can think of the electron as moving not into empty space but into the "graphene vacuum" of virtual electron-hole pairs and thus oscillating back and forth into the hole component of these virtual pairs created around as the particle moves. This is very much like a photon which moves by alternatively going from electric field (A) to magnetic field excitation (B) . The analogy with relativistic quantum mechanics now also follows by noting that in relativity, the energy of a particle is given by (p-momentum):

$$E = \left[(mc)^2 + (pc)^2\right]^{1/2} \tag{5.72}$$

so that the linear dispersion follows in the limit of zero mass.

Remember from Chap. 4 that in the Dirac equation, *the spin does not disappear in the nonrelativistic limit* but remains a fundamental property of quantum particles in a four-dimensional quantum space time. In other words, even when the mean particle velocity is slow, the short-time (light speed) visits into the antiparticle space which is the origin of the spin, are always allowed by Heisenberg's uncertainty principle. Now one can understand why the zero gap nature of semiconductors, graphene being an example, can be so exciting. The visits into hyperspace are now zero energy visits into the valence holes and back (A, B sublattices) and giving rise to a new pseudo spin quantum number. If we now generate a gap, then this can change the quantum dynamics and properties drastically. This is in principle relatively straightforward to produce by external means (gate, multilayer, doping, etc.). Finally, we note that whereas parabolic electrons have constant density of state in 2D, graphene electrons will have linearly increasing density of states with energy (Fig. 5.21).

The short section here does not do justice to the enormously interesting field. The reader is advised to consult the excellent reviews in the literature and in particular see, for example, the excellent review by Phaedon Avouris "Graphene Electronic and Photonic Properties and Devices" Nanoletters vol 10, p4285, (2010).

5.5.3 Two-Dimensional Metal-Dichalcogenide TMDC: Electronic Structures

Introduction

After Graphene, researchers tried to find new types of graphene like 2D layered materials in order may be to discover some of the exciting photonic like electron band structures. Various groups around the world have found out how to make

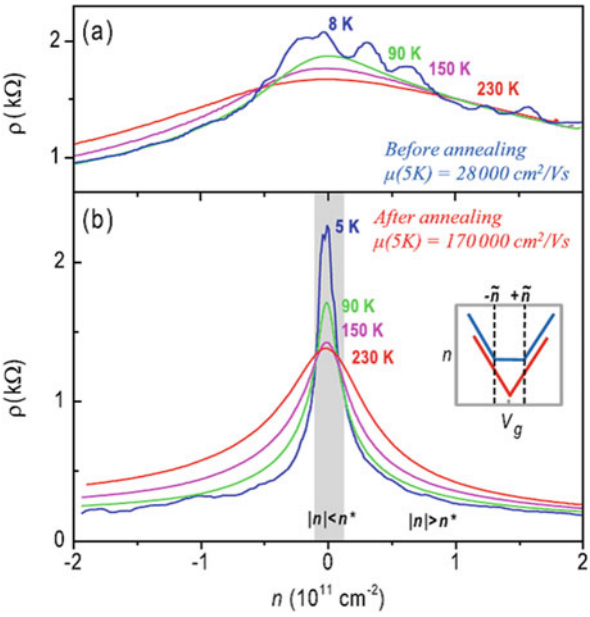


Fig. 5.21 From "Temperature-Dependent Transport in Suspended Graphene" *Bolotin et al.* (2008). "Temperature dependence of resistance of suspended graphene device before and after current annealing. Inset sketch of gate voltage dependence of the carrier density in clean and charge inhomogeneous graphene"

freestanding barrier layers like h-BN, and then finally more recently, they discovered how to exfoliate 2D layers of the transition metal dichalcogenides TMDs [Manish Chewola] . Artificial multilayer fabrication technology is now a very active and popular field of science and technology. Notable discoveries are the ultrathin film transparent high ratio field-effect transistors TFT which exhibit a high degree of plasticity and are promising for tattoo electronics and many other highly commercial applications. Examples of TMD are given in Table 5.1 (Fig. 5.22).

Making nanosheets: description

The TMD sheets have so far been made in several ways described in the work of Chhowalla et al.

1. Scotch tape exfoliation
2. Liquid exfoliation with selected surfactant with right surface energy penetrating the layers and dissolving them
3. Chemical vapor deposition CVD (Figs. 5.23 and 5.24)

5.5.4 Example: Fabrication of Flexible Transistors

Ten atomic thick high-mobility transparent TFTs with ambipolar device characteristics fabricated on both conventional silicon platform and on a flexible substrate have been demonstrated by Saptarshi Das et al. *Nano Letters Vol.14,*

Table 5.1 Electronic character of different layered TMDs[25]

Group	M	X	Properties
4	Ti, Hf, Zr	S, Se, Te	Semiconducting ($E_g = 0.2 \sim 2$ eV). Diamagnetic
5	V, Nb, Ta	S, Se, Te	Narrow band metals ($\rho \sim 10^{-4}$ Ω.cm) or semimetals. Superconducting. Charge density wave (CDW). Paramagnetic, antiferromagnetic, or diamagnetic
6	Mo, W	S, Se, Te	Sulfides and selenides are semiconducting ($E_g \sim 1$ eV). Tellurides are semimetallic ($\rho \sim 10^{-3}$ Ω cm). Diamagnetic
7	Tc, Re	S, Se, Te	Small-gap semiconductors. Diamagnetic
10	Pd, Pt	S, Se, Te	Sulfides and selenides are semiconducting ($E_g = 0.4$ eV) and diamagnetic. Tellurides are metallic and paramagnetic. $PdTe_2$ is superconducting

ρ, in-plane electrical resistivity

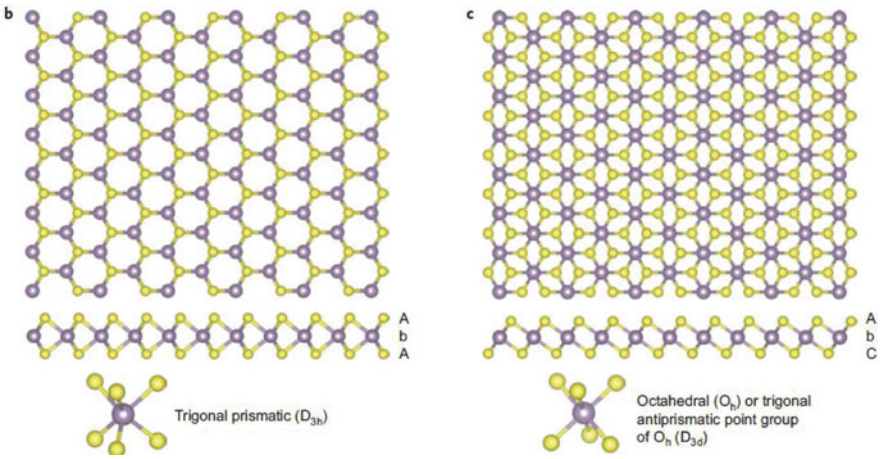


Fig. 5.22 Structure of monolayered TMD. About 40 different layered TMD compounds exist. The transition metals and three chalcogen elements predominantly crystallize in those layered structures. From Chhowalla et al. (2013) [9]

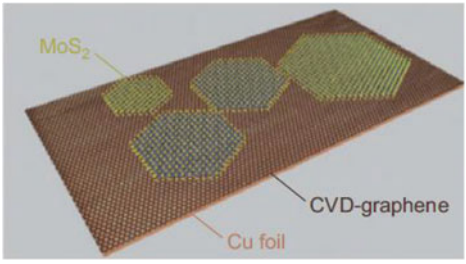


Fig. 5.23 Illustration of the quality achievable for heterostructures MoS_2 on graphene

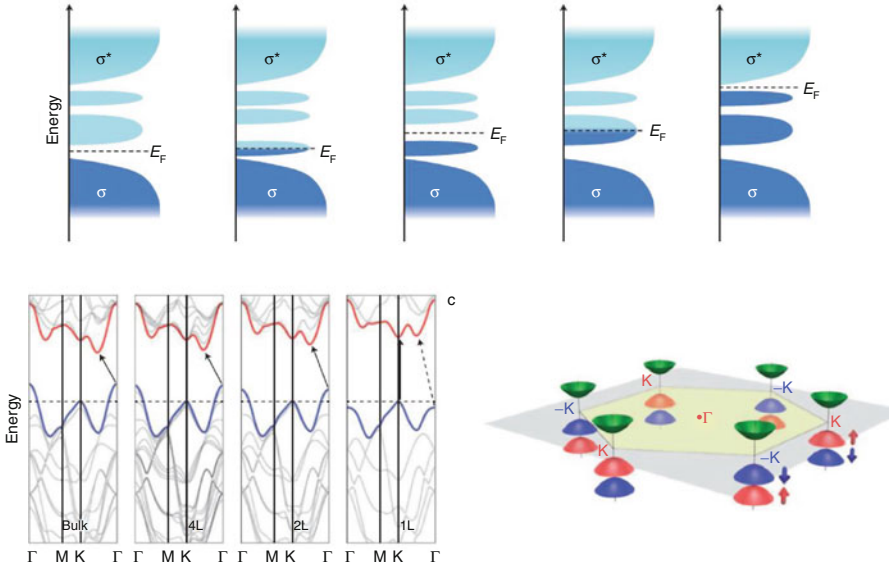


Fig. 5.24 *Qualitative schematic illustration* showing the progressive filling of d-orbitals that are located within the bandgap of bonding and antibonding in groups 4,5,6,7, and 10 TMDs. The D_{3h} and D_{3d} refer to the point groups associated with the trigonal prismatic and the octahedral coordination of the transition metal oxides. (From Chhowalla et al. (2013))

p. 2861, (2014). Monolayer graphene was used as gate electrode and 3–4 atomic layers thick h-BN was used as the gate dielectric, and finally bilayers of WSe_2 were used as the semiconducting channel material for the TFT. The active device stack was found to be 88% transparent over the entire visible spectrum. On to off ratios of 10^7 were observed in all the two-dimensional TFTs.

5.5.5 Summary: Discussion

The atomically thin 2D nanosheets of TMD derived from layered materials exhibit excellent electronic properties, exceptional mechanical flexibility, and partial optical transparency. Beyond the typical semiconducting properties, the various 2D layered materials can also exhibit superconductivity ($NbSe_2$) magnetic ($CrSe_2$) insulating (InN) and thermoelectric (Bi_2Te_3). The 2D sheets can be grown on top of each other to build superlattices with van der Waals bonded layers with exciting new prospects; see the excellent review by Xidong Duan et al.

Xidong Duan et al. review Chem Soc Rev. Vol. 44 p. 8859 (2015): "Two-dimensional transition metal dichalcogenides as atomically thin semiconductors: opportunities and challenges."

Previously in this chapter, we investigated the new "wonder material" called graphene. We described its band structure and explained why this two-dimensional

perfect semimetal with high electron mobility is expected to, and indeed shows, new physical properties which can have serious technical applications. There is by now a massive literature on this subject; indeed, both G and TMD sheets and the interested reader are encouraged to consult some of this material.

5.6 Band Structures in Metals

Although this chapter was primarily devoted to the band structures of semiconductors, which is of great importance in solid-state devices, it would not be complete without a few words on the band structures of metals. Figures 5.25 and 5.26 are examples of electron band structures of two such metals, aluminum and copper.

As mentioned earlier in this chapter, very different behaviors can be seen between the band structures of metals and semiconductors. First of all, there is no forbidden energy region (bandgap) in metals. All the energy range drawn in these diagrams is allowed in metals, which is the most critical difference between metals and semiconductors. Even at a temperature of zero K, a metal has a band which is partially filled with electrons and its Fermi level thus lies within this band. There is no such distinction as valence and conduction bands as encountered in a semiconductor.

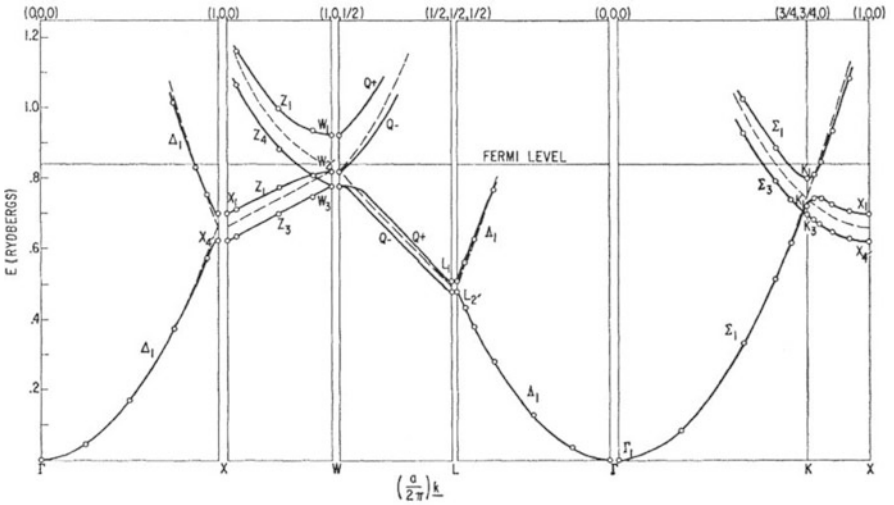


Fig. 5.25 Electron band structure diagram of aluminum. The energy is expressed in units of Rydberg. The dashed lines show the energy bands for a free electron (Reprinted figure with permission from Segall B The Physical Review, vol. 124, p. 1801, Fig. 3, Copyright 1961 by the American Physical Society)

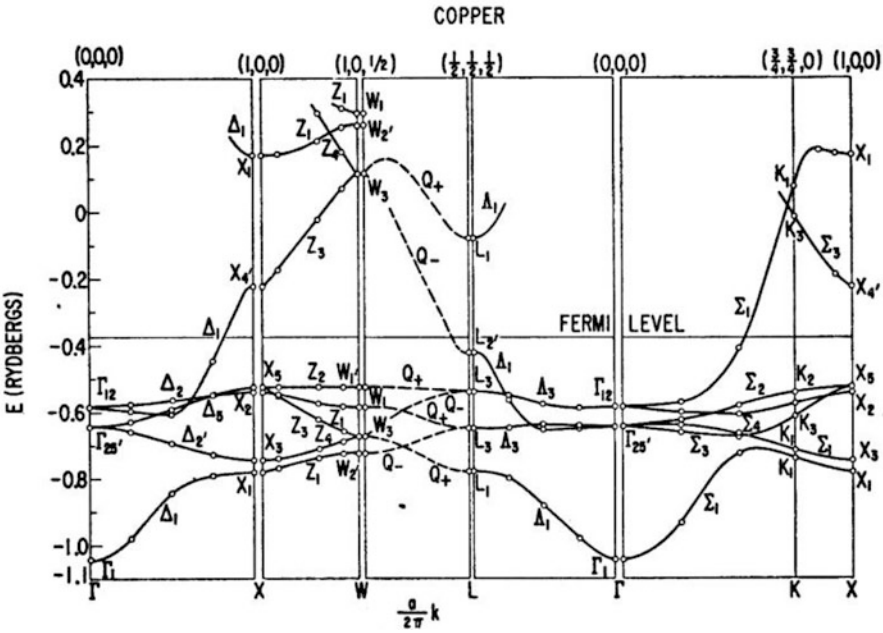


Fig. 5.26 Electron band structure diagram of copper. The energy is expressed in units of Rydberg. There are a few narrow bands located just below the Fermi energy, corresponding to the 4d orbitals in copper (Reprinted figure with permission from Segall B The Physical Review, vol. 125, p. 113, Fig. 5, Copyright 1962 by the American Physical Society)

The band structures in the Γ→X, Γ→K, and Γ→L directions are nearly parabolic and are therefore similar to the free electron case. Electrons in aluminum thus behave almost like free electrons.

The dashed lines in Figs. 5.25 and 5.26 are the *E*–*k* relation for a free electron. One can see that the band structure in aluminum is very close to that of free electrons. The energy spectrum of copper has less resemblance to the free electron *E*–*k* parabolic relation. The major difference between copper and aluminum is the presence of a number of narrow bands below E_F in copper. These narrow bands are attributed to the 4*d*-orbitals of copper atoms. The presence of these *d*-orbital-originated bands is a common feature of most transition metals (such as iron and nickel) and noble metals (such as copper, gold, and silver). These provide a degree of screening effect for electrons. The absence or presence of these *d*-band electrons is also at the origin of the gray and red color appearance of aluminum and copper, respectively. Indeed, when there is a *d*-band, as in copper, not all the photons reaching the metal surface are reflected, but those photons with sufficient energy can be absorbed by the *d*-electrons (see Chapter 0). As a result of this "deficiency" of photons with certain energies, the copper appears red. A similar explanation is valid for the yellow color of gold.

There are always many nearly free electrons in metals that contribute to the electrical and thermal conduction. On the contrary, semiconductors do not have many free electrons when they are intrinsic (i.e., without impurities), and carriers must be provided by a process called doping. The controllability of the doping level in semiconductors is one of the most important reasons why semiconductors are useful in making electronic and optoelectronic devices and will be discussed later in this textbook.

5.7 The Kane Effective Mass Method

In the chapter on band structure, we made the observation, and indeed used this later also throughout the book, that the band dispersion in the majority of semiconductors near $\boldsymbol{k} = 0$ could be approximated as a parabola in $\boldsymbol{k}$ but with an effective mass which is determined by rigorous band structure computation. One finds in practice that the scheme works very well and that the true effective masses can be very different from the free electron masses. From the "exact results" shown in this chapter, one cannot easily understand why the effective mass behaves in the way it does, and one cannot see how it would correlate with the other features of the material, such as its band gap, for example. Also it would be nice to have a scheme which could predict the effective mass, was versatile, and could be applied to confined and multilayer structures as well. Some years ago, Evan O. Kane discovered that it was possible, with rather simple mathematical methods, to shed light on this question. He worked out a scheme with which it is possible to obtain a good approximation to the effective mass near the $\boldsymbol{k} = \mathbf{0}$ points in semiconductors, and a correlation between the effective mass and the band gap.

Kane's method is a brilliant example on how one "piece of information," normally obtained by experiment, can be used to derive another piece of information using the logical structure of a theory. The Kane argument goes as follows.

Consider the full Hamiltonian and Schrödinger equation (SE) of the electron in the periodic potential $V(\vec{r})$ of the lattice. Now assume that the wavefunction is a Bloch wave and must mathematically have the structure:

$$\psi_{n\vec{k}}(\vec{r}) = u_{n\vec{k}}(\vec{r})e^{i\vec{k}.\vec{r}} \tag{5.62}$$

with energy $E_n(\vec{k})$ we know that this must be true, so we substitute it in the SE

$$\left[\frac{p^2}{2m_0} + V(\vec{r})\right]\Psi_{n\vec{k}}(\vec{r}) = E_{n\vec{k}}\Psi_{n\vec{k}}(\vec{r}) \tag{5.63}$$

differentiate, collect the terms, and find

$$\left[\frac{p^2}{2m_0}+\frac{\hbar^2}{m_0}\vec{k}\cdot\vec{p}+V(\vec{r})\right]u_{\vec{nk}}(\vec{r})=\left[E_{\vec{nk}}-\frac{\hbar^2}{2m_0}k^2\right]u_{\vec{nk}}(\vec{r}) \tag{5.64}$$

This is now an equation for the unknown modulating part of the wavefunction $u_{\vec{nk}}(\vec{r})$. The known part has been incorporated and has given an energy shift and a new term in the Hamiltonian. We can rewrite Eq. (5.64) as:

$$\left[H_0+\frac{\hbar^2}{m_0}\vec{k}\cdot\vec{p}\right]u_{\vec{nk}}(\vec{r})=\left[E_{\vec{nk}}-\frac{\hbar^2}{2m_0}k^2\right]u_{\vec{nk}}(\vec{r}) \tag{5.65}$$

and taking the limit $\boldsymbol{k}=\mathbf{0}$, we have the eigenvalue equation:

$$H_0u_n(\vec{r})=[E_n(0)]u_n(\vec{r}) \tag{5.66}$$

for the $\boldsymbol{k}=\mathbf{0}$ envelope. So now one can ask what is the gain in all this, since we are back at the usual Schrödinger equation for the band? There are two observations to be made: the wavefunctions $u_n(\vec{r})$ only have band indices n, there are as many of them as we have energy bands in the semiconductors. In particular, there are valence band functions and conduction band functions. There is a finite energy difference between each band. We could use these functions even though we do not know them, as basis functions, and expand the $\boldsymbol{k}$-dependent term of the Hamiltonian Eq. (5.65) as a perturbation near $\boldsymbol{k}=0$. In this way, we derive the additional $\mathbf{k}$-dependence of the energy and the $\boldsymbol{k}$-dependence of the core wavefunction $u_{\vec{nk}}(\vec{r})$. In this way, we also automatically get an expression for the effective mass in terms of the matrix elements of these basis functions and the energy difference. Thus applying second-order perturbation theory from Chap. 4 to the $\boldsymbol{k}$-dependent term in Eq. (5.65), we have for the energy and wavefunction:

$$E_n(\vec{k})=E_n(0)+\frac{\hbar^2k^2}{2m_0}+\frac{\hbar}{m_0}\vec{k}\cdot\vec{p}_{nn}+\frac{\hbar^2}{m_0^2}\sum_{n'\neq n}\frac{\left|\vec{k}\cdot\vec{p}_{nn'}\right|^2}{E_n(0)-E_{n'}(0)} \tag{5.67}$$

$$u_{\vec{nk}}(\vec{r})=u_{n0}(\vec{r})+\sum_{n'\neq n}\left[\frac{\hbar}{m_0}\frac{\vec{k}\cdot\vec{p}_{n'n}}{E_n(0)-E_{n'}(0)}\right]u_{n'0}(\vec{r}) \tag{5.68}$$

Remember that the complete wavefunction is of the form of Eq. (5.62). Now, we see that progress has indeed been made. When we look at Eq. (5.67), then indeed Eq. (5.67) with $E_{\vec{nk}}\sim E_n(0)+\frac{\hbar^2k^2}{2m^*}$ and the observation that by symmetry $\vec{p}_{nn}=0$ tells us that the effective mass near $\boldsymbol{k}=0$ is given by (i, j denote the x,y,z components)):

$$\left(\frac{1}{m^*}\right)_{ij} = \frac{1}{m_0}\delta_{ij} + \sum_{n' \neq n} \frac{p^i_{nn'} p^j_{n'n} + p^j_{nn'} p^i_{n'n}}{E_n(0) - E_{n'}(0)} \tag{5.69}$$

The inverse of the effective mass is a sum of the free electron mass and term which depends on the momentum matrix elements of the $\boldsymbol{k} = \boldsymbol{0}$ envelope but is also dependent on the energy difference between the bands. If for simplicity, we now consider just two bands, namely, the conduction and valence band, then to a good approximation, we see that the inverse effective mass scales as the inverse of the energy gap of the semiconductor. In other words, we have the result that semiconductors with smaller band gaps should have the lower effective mass. If this statement turns out to be generally true, then it helps to establish an important principle and correlation between band gap and effective mass (see the data in Appendix A4).

At this stage, the most important unknown is the momentum matrix element. The next step is therefore to establish empirically that the momentum matrix elements are not strongly dependent on the band gap and to include the other bands when necessary. Here one also uses the fact that the exact wavefunctions are *s*-like near the bottom of the conduction band and *p*-like near the top of the valence band. This is known from first principle and tight-binding band structure theories. This interplay between theory and experiment then gives us useful and simple empirical rules and numbers for the above matrix element in Eq. (5.69). For example, one finds that the Kane parameter $E_{\mathrm{P}} = \frac{2m_0}{\hbar^2}P^2$ where $P = \frac{\hbar}{m_0}p^z_{cv}$ is roughly 20–25 eV for most semiconductors of interest, where the subscripts c and v denote the conduction and valence band, respectively. The Kane method of expanding around the $\boldsymbol{k} = \boldsymbol{0}$ envelope states can be extended to treat also the spin-orbit interaction. The spin-orbit coupling is of the form:

$$V_{so} = \frac{\hbar}{4m_0^2c^2}\,\vec{\sigma}\cdot\left(\nabla V\left(\vec{r}\right)\times\vec{p}\right) \tag{5.70}$$

where σ is the electron spin operator and $V\left(\vec{r}\right)$ is the total potential experienced by the electrons. The spin-orbit interaction is a small but non-negligible effect in semiconductors. It is ideally treated using the Kane model because the energy shifts up to second order in perturbation theory and involved the same type of matrix elements of the momentum as before. Indeed, one can say that the Kane method provides a very natural way to treat the spin-orbit interaction. The method can be extended to also treat confined systems. The results can at the end be expressed as functions of E_{g} and P. The first of which, E_{g}, is known, and the second of which, P, can be estimated to good accuracy.

Kane theory tells us that the effective mass is related to the structure of the envelope momentum matrix elements. These as it happens do not change all that much from one system to another system. The band gap which also enters the formula, however, changes quite a lot. If for some reason, such as strain or

confinement, the band gap changes, even locally, then we can expect the effective mass also to change locally. The changes in P or wavefunction shapes are of lower order than the band gap changes, and this is why the Kane method is so useful. The Kane method is therefore a very practical way of handling strain effects in semiconductor interfaces. This happens when there is lattice mismatch forcing the top grown lattice to adopt the lattice parameters of the substrate. The mismatch can force the top layer bonds to be stretched or compressed. Compression or dilation affects both Kane parameters E_g and P locally. But the gap is more sensitive than P to first order. In quantum dots strain, one also has strain which can vary locally and give rise to local effective mass. The reader is referred to the book by L Chuang for a detailed treatment of the Kane model and its applications.

5.7.1 The Effect of the Spin-Orbit Coupling

Let us now consider the effect of the spin-orbit coupling explicitly. Let us go back to Eq. (5.63) and include the spin-orbit interaction:

$$\left[\frac{p^2}{2m_0}+V(\vec{r})+\frac{\hbar}{4m_0^2c^2}\left[\vec{\nabla}\ V\times\ \vec{p}\right]\cdot\ \vec{\sigma}\right]\Psi_{n\vec{k}}(\vec{r})=E_{n\vec{k}}\Psi_{n\vec{k}}(\vec{r}) \tag{5.71}$$

Substituting the Bloch function then gives:

$$\left\{\frac{p^2}{2m_0}+\frac{\hbar}{m_0}\ \vec{k}\cdot\vec{p}\ +V(\vec{r})+\frac{\hbar}{4m_0^2c^2}\left[\vec{\nabla}\ V\times\ \vec{p}\right]\cdot\ \vec{\sigma}\right.$$
$$\left.+\frac{\hbar^2}{4m_0^2c^2}\left[\nabla V\times\ \vec{k}\ \right]\cdot\ \vec{\sigma}\ \right]u_{n\vec{k}}(\vec{r})\Bigg\}=\left[E_{n\vec{k}}-\frac{\hbar^2}{2m_0}k^2\right]u_{n\vec{k}}(\vec{r}) \tag{5.72}$$

Following the book by L. Chuang (see references list), we will define:

$$E^{'}=E-\frac{\hbar^2k^2}{2m_0} \tag{5.73}$$

The last term in Eq. (5.72) depends on the Bloch wavevector $\boldsymbol{k}$ and is much smaller than the term which involves the momentum operator. The reason is that the momentum around the nucleus is much larger than the band momentum $\boldsymbol{k}$ so that we can neglect the last term to obtain

$$\left\{\frac{p^2}{2m_0}+\frac{\hbar^2}{m_0}\ \vec{k}\ .\ \vec{p}\ +V(\vec{r})+\frac{\hbar}{4m_0^2c^2}\left[\vec{\nabla}\ V\times\ \vec{p}\right]\cdot\ \vec{\sigma}\ \right]\Big\}u_{n\vec{k}}(\vec{r})\Big\}=\left[E^{'}\right]u_{n\vec{k}}(\vec{r}) \tag{5.74}$$

In order to solve this equation in the $\boldsymbol{k}\cdot\boldsymbol{p}$ approach, we assume as before that the wavefunction can be written as a superposition of the $\boldsymbol{k}=0$ subbands:

$$u_{n\vec{k}} = \sum_{n'} a_{n'\vec{k}} u_{n'0}(r) \tag{5.75}$$

The $u_{n'0}(\vec{r})$ are chosen to roughly correspond to what one knows about the system from first principle band structure techniques, namely, that for the eigenstate near the conduction band edge, the wavefunctions have S symmetry. Those near the top of the valence band have p-symmetry so that

$$\text{Conduction bands} |S\uparrow\rangle, |S\downarrow\rangle$$

$$\text{Valence bands} |X\uparrow\rangle, |Y\uparrow\rangle, |Z\uparrow\rangle, |X\downarrow\rangle, |Y\downarrow\rangle, |Z\downarrow\rangle$$

Also these wavefunctions satisfy the conditions $H_0|S\uparrow\rangle = E_S|S\uparrow\rangle$ without a magnetic field, and the two spin states have the same energy, and $H_0|X\uparrow\rangle = E_p|X\uparrow\rangle$, $H_0|Y\uparrow\rangle = E_p|Y\uparrow\rangle$, and $H_0|Z\uparrow\rangle = E_p|Z\uparrow\rangle$. For practical purposes, it is convenient to choose the basis states for spin and angular momentum raising and lowering operators:

$$|S\uparrow\rangle, \left|\frac{1}{\sqrt{2}}(X - iY)\uparrow\right\rangle, |Z\downarrow\rangle, \left|-\frac{1}{\sqrt{2}}(X + iY)\uparrow\right\rangle \tag{5.76}$$

$$|S\downarrow\rangle, \left|-\frac{1}{\sqrt{2}}(X + iY)\downarrow\right\rangle, |Z\uparrow\rangle, \left|\frac{1}{\sqrt{2}}(X - iY)\downarrow\right\rangle$$

The valence band basis states can be selected from the eigenstates of angular momentum $\boldsymbol{L}$ in Chap.4. So a p-state corresponds to $l = 1$ and we have:

$$Y_{1,\pm 1} = \mp\frac{1}{\sqrt{2}}|X \pm iY\rangle \quad \text{and} \quad Y_{10} = |Z\rangle \tag{5.77}$$

Now we can generate the matrix representation of the Hamiltonian Eq. (5.74) using this basis set to find an 8×8 matrix which as a result of spin degeneracy reduces to a 4×4 matrix:

$$\begin{pmatrix} E_s & 0 & kP & 0 \\ 0 & E_p - \frac{\Delta}{3} & \frac{\sqrt{2}}{3}\Delta & 0 \\ kP & \frac{\sqrt{2}\Delta}{3} & E_p & 0 \\ 0 & 0 & 0 & E_p + \frac{\Delta}{3} \end{pmatrix} \tag{5.78}$$

where the Kane parameter is defined as:

$$\begin{aligned} P &= -i\frac{\hbar}{m_0}\langle S|p_z|Z\rangle \\ \Delta &= \frac{3\hbar i}{4m_0^2c^2}\left\{\langle X|\frac{\partial V}{\partial x}p_y - \frac{\partial V}{\partial y}p_x|Y\rangle\right\} \end{aligned} \tag{5.79}$$

Let us measure the eigenvalues of this system such that the conduction band E_s is at E_g and the top of the valence band is at 0. The solutions are the roots of the equation:

$$E'\left(E' - E_g\right)\left(E' + \Delta\right) - k^2P^2\left(E' + \frac{2}{3}\Delta\right) = 0 \tag{5.80}$$

which we can solve analytically if we expand to first order in k^2. The result, for the energy and effective mass in the conduction band, is:

$$E_c(k) = E_g + \frac{\hbar^2k^2}{2m_0} + \frac{k^2P^2}{3}\frac{\left(3E_g + 2\Delta\right)}{E_g\left(E_g + \Delta\right)} \tag{5.81}$$

$$\frac{1}{m_c^*} = \frac{1}{m_0} + \frac{2P^2}{3\hbar^2}\frac{\left(3E_g + 2\Delta\right)}{E_g\left(E_g + \Delta\right)} \tag{5.82}$$

For the heavy-hole valence states, we have:

$$E_{hh}(k) = \frac{\hbar^2k^2}{2m_0} \tag{5.83}$$

$$\frac{1}{m_{hh}^*} = \frac{1}{m_0} \tag{5.84}$$

For the light hole:

$$E_{lh}(k) = \frac{\hbar^2k^2}{2m_0} - \frac{2k^2P^2}{3}\frac{1}{E_g} \tag{5.85}$$

$$\frac{1}{m_{lh}^*} = \frac{1}{m_0} - \frac{4P^2}{3\hbar^2}\frac{1}{E_g} \tag{5.86}$$

For the spin orbit shifted band:

$$E_{so}(k) = -\Delta + \frac{\hbar^2k^2}{2m_0} - \frac{k^2P^2}{3}\frac{1}{\left(E_g + \Delta\right)} \tag{5.87}$$

$$\frac{1}{m_{so}^*} = \frac{1}{m_0} - \frac{2P^2}{3\hbar^2}\frac{1}{\left(E_g + \Delta\right)} \tag{5.88}$$

To zero order in k^2, the conduction band wavefunctions are unchanged at $|S\uparrow\rangle$ and $|S\downarrow\rangle$. The valence light-hole states have a spin-orbit shift even to this order. So we have for the heavy hole the two states:

$$\left|-\frac{1}{\sqrt{2}}(X+iY)\uparrow\right\rangle. = \left|\frac{3}{2},\frac{3}{2}\right\rangle. \tag{5.89}$$

$$\left|\frac{1}{\sqrt{2}}(X-iY)\downarrow\right\rangle. = \left|\frac{3}{2},-\frac{3}{2}\right\rangle. \tag{5.90}$$

and the light holes:

$$\left|\frac{1}{\sqrt{6}}(X-iY)\uparrow\right\rangle. + \sqrt{\frac{2}{3}}|Z\downarrow\rangle. = \left|\frac{3}{2},-\frac{1}{2}\right\rangle. \tag{5.91}$$

$$-\left|\frac{1}{\sqrt{6}}(X+iY)\downarrow\right\rangle. + \sqrt{\frac{2}{3}}|Z\uparrow\rangle. = \left|\frac{3}{2},\frac{1}{2}\right\rangle. \tag{5.92}$$

The $\boldsymbol{k}\cdot\boldsymbol{p}$ perturbation result can be generated as in Eqs. (5.67) and (5.68). The exact result can be obtained by solving for the eigenvalues $E_n^{'} = E_n - \dfrac{\hbar^2 k^2}{2m_0}$, as done above, then substituting back to solve the linear equations. The matrix is effectively 3×3 so that:

$$\begin{pmatrix} E_g - E_n' & 0 & kP \\ 0 & -\dfrac{2\Delta}{3} - E_n^{'} & \dfrac{\sqrt{2}}{3}\Delta \\ kP & \dfrac{\sqrt{2}\Delta}{3} & -\dfrac{\Delta}{3} - E_n^{'} \end{pmatrix} \begin{pmatrix} a_n \\ b_n \\ c_n \end{pmatrix} = 0 \tag{5.93}$$

with

$$\left\{|a_n|^2 + |b_n|^2 + |c_n|^2\right\}^{1/2} = 1 \tag{5.94}$$

Note that the present approach neglects the remote band effects, and it does not reproduce the correct heavy-hole mass. In order to do that, one has to go further and consider the Luttinger-Kohn model which is similar in spirit but takes into account remote energy bands and will not be considered here.

The k-dependence of the wavefunctions has not been studied here. They can be of great interest in doped magnetic semiconductors and magnetic metals where in the presence of a finite spin polarization, they can give rise to the so-called anomalous Hall effect and spin Hall effect (see Jungwirth et al. 2006). But again here, one would go further and use the Luttinger-Kohn model which includes the remote band effects. The magnetism can then be treated as an effective uniform self-consistent Curie field which acts on the spin system (see Jungwirth et al. reference).

5.7.2 Summary

In this chapter, using simple quantum mechanical concepts and methods, we have described the energy states of electrons in a periodic potential. We have modeled the crystal using the Kronig-Penney model. Nearly free electron and the tight-binding approximations were briefly introduced. We familiarized the reader with the notion of band structure, band gap, Bloch wavefunction, effective mass, Fermi energy, and Fermi-Dirac distribution and holes. The band structures for the common semiconductors, including Si, Ge, and GaAs, have been illustrated after first describing the conventionally used high-symmetry points and orientations. The main features in these band structures have been outlined. The band structures of a few metals, including aluminum and copper, have also been presented, and the main features were described and compared to those of semiconductors. We have shown how one can evaluate the Bloch wavefunctions and effective masses of semiconductors near $k = 0$ using a scheme called the $k{\cdot}p$ method. The method is simple and very powerful. It was applied to derive the effective mass including the spin-orbit coupling.

Problems

1. *Equations of motion of an electron in the presence of an electric field.*

$$\text{Assuming a dispersion relation : } \varepsilon = \varepsilon_C + \frac{\hbar^2}{ma^2}[1 - \cos(ka)]$$

   (a) Calculate the velocity of the electron at $k = \pi/a$.
   (b) If the electric field E is applied in the $-x$ direction, derive the time dependence of k for an electron initially at $k = \pi/a$ and position $x = 0$.
   (c) Derive the time dependence of the electron velocity, $v(t)$, and the time dependence of the electron position, $x(t)$.
   (d) For $a = 5$ nm, $E = 104$ V·cm^{-1}, and $m = 0.2\ m_0$, what are the maximum and minimum values of x that the electron will reach?
   (e) What is the period of the oscillation?
   (f) For the parameter of part (e), derive an expression for the effective mass as a function of k. Sketch the function.

2. *The period of the Bloch oscillations.*

   Consider an electron that is subjected to an electric field. The electric field exerts a force $F = -qE$ on the electron. Assume that the electron is initially not in motion, i.e., $k = 0$. Upon application of the electric field, the k value of the electron increases from 0 to π/a. At this value of k, Bragg reflection occurs, and the electron assumes a k value of $-\pi/a$. Then, the electron is again accelerated to $k = \pi/a$. At this point, the electron again undergoes Bragg reflection, and the cycle starts from the beginning. The process described above is called the Bloch oscillation of the electron in an energy band of the solid-state crystal.

(a) Show that the period of the Bloch oscillation is given by $\tau = \frac{2\pi\hbar}{qEa}$, where a is the periodicity of a one-dimensional atomic chain.

(b) Calculate the period of the Bloch oscillations for $a = 4$ Å and $E = 1250$ V·cm^{-1}. Compare the period of the Bloch oscillations with a typical inelastic scattering times. What conclusions do you draw from the comparison? Are the Bragg reflections important scattering events for the movement of electrons in a crystal? Typical inelastic scattering times are 10^{-11} s for low fields and 10^{-13} s for high fields.

3. *Idealized electron dynamics.*

A single electron is placed at $k = 0$ in an otherwise empty band of a bcc solid. The energy versus k relation of the band is given by:

$$\varepsilon\left(\vec{k}\right) = -\alpha - 8\gamma \cos\left(\frac{k_x a}{2}\right).$$

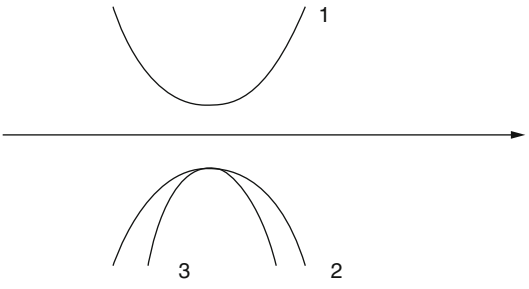
At $t = 0$, a uniform electric field E is applied in the x-axis direction. Describe the motion of the electron in k-space. Use a reduced-zone picture. Discuss the motion of the electron in real space assuming that the particle starts its journey at the origin at $t = 0$. Using the reduced-zone picture, describe the movement of the electron in k-space. Discuss the motion of the electron in real space assuming that the particle starts its movement at the origin at $t = 0$.

4. *Effective mass.*

For some materials, the band structure of the conduction band around $k = 0$ can be represented by $\varepsilon\left(\vec{k}\right) = \frac{\hbar^2}{2m} A\left(k_x^2 - \frac{a^2}{2\pi^2}k_x^4\right)$.

What is the effective mass of a free electron under these conditions?

On the figure, name the different bands and point out which one of the two in the lower band has the higher effective mass.



5. *Calculate the coordinates of the high-symmetry point U in Fig. 5.15.*

6. *Origin of electronic bands in materials.*

Explain how electronic energy bands arise in materials.

The periodic potential in a one-dimensional lattice of spacing a can be approximated by a square wave which has the value $U_0 = -2$ eV at each

atom and which changes to zero at a distance of $0.1a$ on either side of each atom. Describe how you would estimate the width of the first energy gap in the electron energy spectrum.

7. *Position of the Fermi level in intrinsic semiconductors.*
   Assume that the density of states is the same in the conduction band (N_C) and in the valence band (N_V). Then, the probability p that a state is filled at the conduction band edge (E_C) is equal to the probability p that a state is empty in the valence band edge (E_V). Where is the Fermi level located?
8. *Plot of the Fermi distribution function at two different temperatures.*
   Calculate the Fermi function at 6.5 eV if $E_F = 6.25$ eV and $T = 300$ K. Repeat for $T = 950$ K assuming that the Fermi energy does not change. Plot the energy dependence of the electron distribution function at $T = 300$ K and at $T = 950$ K assuming $E_F = 6.25$ eV.
9. *Numerical evaluation of the effective densities of states of Ge, Si, and GaAs.*
   Calculate the effective densities of states in the conduction and valence bands of germanium, silicon and gallium arsenide at 300 K. Note in analogy to Eq. (5.55) we have $N_V = 2\left(\frac{2\pi k_b T m_h}{h^2}\right)^{3/2}$
10. *Density of states of a piece of Si.*
    Calculate the number of states per unit energy in a 100 by 100 by 10 nm piece of silicon ($m^* = 1.08\ m_0$) 100 meV above the conduction band edge. Write the results in units of eV^{-1}.
11. *Number of conduction electrons in a Fermi sphere of known radius.*
    In a simple cubic quasi-free electron metal, the spherical Fermi surface just touches the first Brillouin zone. Calculate the number of conduction electrons per atom in this metal as a function of the Fermi-Dirac integral. Consider the energy at the bottom of the conduction band to be $E_C = 0$ eV.

References

Chhowalla M et al (2013) Electronics and optolectronics of two –dimensional transition metal dichalcogenides. Nat Chem 5:263
Das S et al (2014) Nano Lett 14:2861
Duan X et al (2015) Review Chem Soc Rev 44:8859
Chelikowsky JR, Cohen ML (1976) Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors. Phys Rev B 14:556–582
Jungwirth T, Sinova J, Masek J, Kucera J, Macdonald AH (2006) Theory of ferromagnetic (III,Mn) V semiconductors. Rev Mod Phys 78:809

Further Reading

Andre G (2011) Nobel lecture: random walk to graphene. Rev Mod Phys 83
Avouris P (2010) Graphene electronic and photonic properties and devices. Nano 10:4285
Bolotin K, Sikes K, Stormer HL, Kim P (2008) PRL 101:096802

Bonaccorso F, Sun Z, Hasan T, Ferrari A (2010) Graphene photonics and optoelectronics. Nat Photonics 4:611

Bonnacorso F et al (2015) Graphene related two-dimensional crystal and hybrid systems for energy conversion and storage. Science 347(6217):1246501–1246501.

Falkovsky L (2008) Optical properties of graphene and IV-VI semiconductors. The electronic properties of graphene. Physics Uspekhi 51:887

Jungwirth T, Sinova J, Masek J, Kucera J, Macdonald AH Theory of ferromagnetic (III,Mn)V semiconductors. Rev Mod Phys 78:809 2006

Neto CAH (2009) Rev Mod Phys 81. Jan–March 2009

Peres NM, Guinea F, Castro AH (2006) Electronic properties of disordered two- dimensional carbon. PRB 73:125411

Wallace PR (1947) Phys Rev 71:622

Phonons and Thermal Properties

6

6.1 Phonons and Thermal Properties

6.1.1 Introduction

In the previous chapters, we have considered the electrons in a crystal that consisted of a rigid lattice of atoms. This represented a good approximation because the mass of an atom is more than 2000 of the mass of an electron. However, such assumptions founder when considering specific heat, thermal expansion, the temperature dependence of electron relaxation time, and thermal conductivity. In order to interpret these phenomena involving electrons and atoms, a more refined model needs to be considered, in which the atoms are allowed to move and vibrate around their equilibrium positions in the lattice. In this chapter, we will present a simple yet relatively accurate mathematical model to describe the mechanical vibrations of atoms in a crystal. We will first cover one-dimensional monatomic and diatomic crystals followed by three-dimensional crystals. We will then consider the collective movement or excitations of the atoms in a crystal, the so-called phonons, and conclude with a section on the velocity of sound in a medium.

6.1.2 Interaction of Atoms in Crystals: Origin and Formalism

We saw in Sect. 1.5, when discussing the formation of bonds in solids, that these equilibrium positions were achieved by balancing attractive and repulsive forces between individual atoms. We assumed that the attractive and repulsive forces always canceled each other and that the masses were infinite. The resulting potential $U(R)$ curve for an atom as a function of its distance R from a neighboring atom is shown in Fig. 6.1. This figure shows a minimum energy for a specific atomic separation, which we understood was true at all time.

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_6

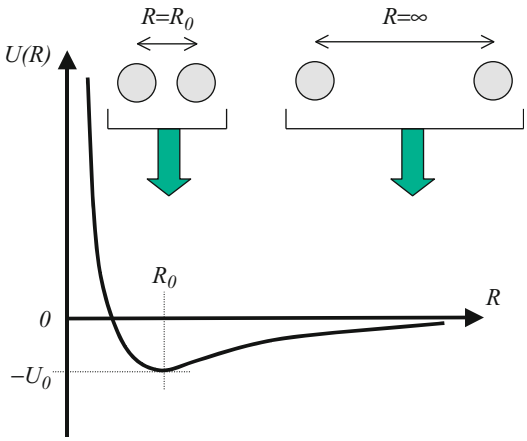


Fig. 6.1 Potential energy of two neighboring atoms in a crystal as a function of the interatomic spacing. When the two atoms are very far away from each other, they do not interact, and the interaction potential energy is near zero. When they get closer to one another, they are attracted to each other to form a bond, which leads to a lowering of the potential energy. However, when they are very close, the electrostatic repulsion from the nuclear charge of each atom leads to a repulsive interaction and an increase in the potential energy

The origin of these forces lies in the electrostatic interaction between the electrical charges (nuclei and electron clouds) in the two neighboring atoms. Classically, the electrons are constantly moving in an atom, in a non-deterministic manner (thus the name "cloud"). One can easily understand that the attractive and repulsive forces do not balance each other at all times but rather the attractive force would be stronger than the repulsive force at a certain time and then weaker shortly afterward. On average, a balance of forces is still achieved. We therefore realize that the positions of atoms in a lattice are not fixed in time but that small deviations do occur around the equilibrium positions. Such vibrations are also more intense at higher temperatures. Note that this is a fully classical analysis of why these lattice vibrations exist. The quantum mechanical description is quite different. In quantum mechanics, the electrons do not move about the lattice in a cloud but occupy energy levels inside allowed energy bands. The lattice atoms have kinetic and potential energy, and the wavefunction for lattice vibrations must also obey Schrödinger equation. The solutions to Schrödinger equation give one the eigenfunctions and allowed energy levels of the lattice vibrations. These allowed energy levels of lattice vibrations are called phonons. In the quantum mechanical description, the lattice is never at rest, even at 0 K. The atoms always move, or oscillate, because the Heisenberg uncertainty principle does not allow the atoms to have a definite position in space. If the atoms were stationary, then their momentum would be indeterminate. The quantum compromise for this scenario is called the zero-point energy which naturally derives from Schrödinger equation and gives the lattice vibrational modes a minimum amount of spatial uncertainty called the zero-point motion. To this zero-point motion, there is a zero-point energy. This observation is already true for the simple

diatomic molecule, for which the vibrational modes are the solutions of the harmonic oscillator problem in quantum mechanics. Instead of solving Schrödinger equation for lattice vibrations, it is much easier and more convenient to first study the allowed classical modes of vibration. It turns out that the classical treatment survives the quantum treatment. The classical bands change into the true quantum lattice energy bands through a simple transformation. We will therefore continue with the more intuitive classical description knowing that the classical results can be taken over in the quantum limit.

Let us now develop a simple mathematical model for such atomic vibrations and introduce the formalism that will be used in the rest of the text. We start by considering the two neighboring atoms, one at the origin ($R = 0$) and the other at a distance R, while its equilibrium position is at $R = R_0$. A one-dimensional analysis will be considered at this time. The potential energy $U(R)$ in Fig. 6.1 of the second atom can be conveniently expressed with respect to the equilibrium values at R_0 through what is called the Taylor expansion (see Appendix A.5):

$$U(R) = U(R_0) + \left(\frac{dU}{dR}\right)_{R_0}(R - R_0) + \frac{1}{2}\left(\frac{d^2U}{dR^2}\right)_{R_0}(R - R_0)^2 + \frac{1}{6}\left(\frac{d^3U}{dR^3}\right)_{R_0}(R - R_0)^3 + \ldots \tag{6.1}$$

where $\left(\frac{dU}{dR}\right)_{R_0}$, $\left(\frac{d^2U}{dR^2}\right)_{R_0}$, $\left(\frac{d^3U}{dR^3}\right)_{R_0}$ are the first, second, and third derivatives of $U(r)$, respectively, evaluated at $r = R_0$. $(R-R_0)$ is called the displacement. The first derivative $\left(\frac{dU}{dR}\right)_{R_0}$ is in fact equal to zero, because it is calculated at the equilibrium position $r = R_0$, which is where the potential $U(r)$ reaches a minimum. Therefore, only the displacement terms $(R-R_0)^n$ with an exponent n larger than or equal to 2 are left. The usefulness of the Taylor expansion resides in the fact that at small deviations from equilibrium, i.e., $(R-R_0) << R_0$, it is reasonable to approximate $U(R)$ with only the first few terms of the expansion in Eq. (6.1).

By denoting $U_0 = -U(R_0)$ and $x = R - R_0$ the displacement, Eq. (6.1) can be rewritten as:

$$U(x) + U_0 = \frac{1}{2}C_1x^2 + C_2x^3 + \ldots \tag{6.2}$$

where $C_1 = \left(\frac{d^2U}{dR^2}\right)_{R_0}$ and $C_2 = \frac{1}{6}\left(\frac{d^3U}{dR^3}\right)_{R_0}$ are constants of the model, determined by the nature of the atoms considered. The first term in the right-hand side of Eq. (6.2), $\frac{1}{2}C_1x^2$, is in fact the potential energy associated with an elastic force equal to $F = -\frac{d}{dx}\left(\frac{1}{2}C_1x^2\right) = -C_1x$, where C_1 is the elastic force constant. The negative sign means that F acts as a restoring force, i.e., in the direction opposite to the displacement u of the atom.

In the following sections, we will limit the analysis to the first term in the expansion in Eq. (6.2) and denote $C = C_1$:

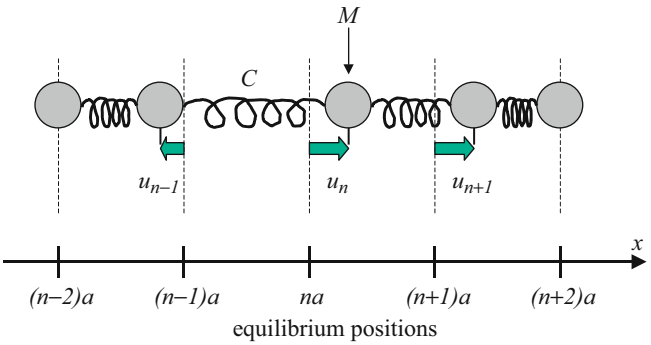


Fig. 6.2 Model for the interaction of identical atoms in a harmonic crystal. The relative movement of the atoms is modeled by a spring such that atoms displaced from their equilibrium positions are forced back by the neighboring atoms. The displacement can travel like a wave throughout the lattice

$$U(x) + U_0 \approx \frac{1}{2}Cx^2 \tag{6.3}$$

Because the atomic vibrations described by this potential only involve second-order displacements, such a solid is generally referred as a harmonic crystal in which the interactions between atoms can be modeled by a spring. This formalism is valid in solids up to all reasonable temperatures. We will apply this formalism to two cases of one-dimensional lattice, extend it to a three-dimensional lattice, and derive a few macroscopic physical properties of crystals.

6.1.3 One-Dimensional Monatomic Harmonic Crystal

In this simple model, we consider a one-dimensional (linear) lattice with a period a and with identical atoms of mass M, vibrating around each lattice point, as depicted in Fig. 6.2. Each atom is indexed by an integer n, and its displacement from its equilibrium position is denoted u_n. The atoms are taken to oscillate in the same direction as the lattice (i.e., longitudinal vibration). All the results obtained for this artificial one-dimensional model prove to be true for three-dimensional lattices as well.

Traveling Wave Formalism

In this one-dimensional case, we will take into account only the interaction between nearest neighbors, an assumption that has little effect on the final results. When considering two neighboring atoms, the forces that are exerted on each one can be

modeled as resulting from a spring which links the interacting atoms, as the one shown in Fig. 6.2. In other words, the force acted on the n^{th} atom:

- By the $(n\text{-}1)^{\text{th}}$ atom is $F_{n,\,n-1} = -C(u_n - u_{n-1})$
- By the $(n+1)^{\text{th}}$ atom is $F_{n,\,n+1} = -C(u_n - u_{n+1})$

where C is the quasi-elastic force constant, a characteristic of the spring. Although this spring formalism is obviously crude, it nevertheless describes the interaction between atoms rather well. This is because the elastic force constant C arises from Eq. (6.3) and corresponds to the first level of approximation for the interactions between atoms. The resultant force acting on the n^{th} atom is therefore:

$$F_n = F_{n,n-1} + F_{n,n+1} = -C(2u_n - u_{n-1} - u_{n+1}) \tag{6.4}$$

The equation of motion for the n^{th} atom is then expressed using classical mechanics like Newton's law:

$$M\frac{d^2u_n}{dt^2} = F_n = -C(2u_n - u_{n-1} - u_{n+1}) \tag{6.5}$$

where M is the mass and $\frac{d^2u_n}{dt^2}$ is the acceleration of the n^{th} atom. We thus obtain a large number of coupled differential equations, where the unknown functions are the displacements $u_n(t)$. We seek solutions to the Eq. (6.5) in the form of traveling waves such as:

$$u_n(t) = A\exp[i(kan - \omega t)] \tag{6.6}$$

where A is the amplitude of the displacement, k is the wavenumber of the wave, and ω its angular frequency. This expression is typical of a traveling wave because it satisfies the relation:

$$\begin{aligned} u_{n+1}(t) &= A\exp[i(ka(n+1) - \omega t)] = A\exp\left[i\left(kan - \omega\left(t - \frac{ka}{\omega}\right)\right)\right] \\ &= u_n\left(t - \frac{ka}{\omega}\right) \end{aligned} \tag{6.7}$$

which shows that the value of the displacement $u_{n+1}(t)$ at the $(n+1)^{\text{th}}$ atom at a time t is the same as the displacement $u_n(t)$ at the n^{th} atom at an earlier time $\left(t - \frac{ka}{\omega}\right)$. This means that the magnitude of the displacement is like a wave that is traveling a distance a in space during a time $\frac{ka}{\omega}$. The velocity at which the wave is traveling is therefore equal to:

$$\frac{a}{ka/\omega} = \frac{\omega}{k} \tag{6.8}$$

The wavelength λ and frequency υ of the traveling wave are related to the wavenumber or angular frequency through the defined relations:

$$\begin{cases} \lambda = \dfrac{2\pi}{k} \\ \upsilon = \dfrac{\omega}{2\pi} \end{cases} \tag{6.9}$$

Boundary Conditions

Before solving the equation of motion in Eq. (6.5), we must introduce the boundary condition that the linear array of atoms is finite and consists of N atoms with the first and last atoms being equivalent, i.e., $u_{n+N}(t) = u_n(t)$. This is the periodic or Born-von Karman boundary conditions which we have already encountered in Sect. 5.3. This is a reasonable assumption because macroscopic crystal specimens consist of a very large number of atoms. And since the interaction forces are significant only between neighboring atoms, the motion of boundary atoms on the "surface" of the specimen does not affect the motion of all other atoms inside the sample.

Because of the general exponential expression of $u_n(t)$ (Eq. (6.6)), these conditions lead to the discretization of the wavenumber k, similar to what was obtained in Chap. 5:

$$k = k_m = \frac{2\pi}{a}\frac{m}{N} \tag{6.10}$$

where $m = 0, \pm 1\ldots$ is an integer. In fact, only N different values of wavenumber k are necessary. Indeed, if two wavenumbers k and k' differ from each other by an integer times $\frac{2\pi}{a}$ (e.g., $k' = k + \frac{2\pi}{a}$), which is equivalent to say that their corresponding integers m and m' differ by an integer times N (e.g., $m' = m + N$), then they lead to the same function $u_n(t)$ as seen through the simple calculation:

$$\begin{aligned} u_n'(t) &= A\exp\left[i\left(k'an - \omega t\right)\right] \\ &= A\exp[i2\pi n + i(kan - \omega t)] = u_n(t) \end{aligned} \tag{6.11}$$

which is valid for any point (na) and any time (t). This means that k and k' are physically indistinguishable. In other words, the basic interval of variation of k can be chosen as:

$$\frac{1}{2}\left(-\frac{2\pi}{a}\right) \le k \le \frac{1}{2}\left(\frac{2\pi}{a}\right) \tag{6.12}$$

And all the physical properties of our one-dimensional crystal that depends on the wavenumber k must be periodic with a period $\frac{2\pi}{a}$. Again, we arrive at the concept of the first Brillouin zone introduced in Chap. 5 and Sect. 5.4.1 for electronic states. And the quantity $\frac{2\pi}{a}$ is a reciprocal lattice period. Of course, we can (and must) always choose the number of atoms N so large that the variation of k could be considered as quasi-continuous.

Phonon Dispersion Relation

Now we can solve the equation of motion in Eq. (6.5), by substitutingEq. (6.6) into it:

$$-M\omega^2 A\exp[i(kan-\omega t)] = -C\left(2-e^{-ika}-e^{ika}\right)A\exp[i(kan-\omega t)]$$

which successively becomes, after simplification of the exponential and the constant A:

$$-M\omega^2 = -C\left(2-e^{-ika}-e^{ika}\right)$$

or:

$$\omega^2 = \frac{2C}{M}(1-\cos ka) = \frac{4C}{M}\sin^2\frac{ka}{2} \tag{6.13}$$

where we have made use of the trigonometric relation: $1-\cos x = 2\sin^2\frac{x}{2}$. This last expression can also be rewritten as:

$$\omega = \omega_{\max}\left|\sin\frac{ka}{2}\right| \tag{6.14}$$

where $\omega_{\max} = \sqrt{\frac{4C}{M}}$. This relation is called the phonon dispersion relation and is plotted in Fig. 6.3.

We see that the solutions of Eq. (6.5) of the traveling wave type exist only if the relation in Eq. (6.14) is satisfied by the wavenumber k and the angular frequency ω of the traveling wave. The frequency and wavenumber of the traveling wave

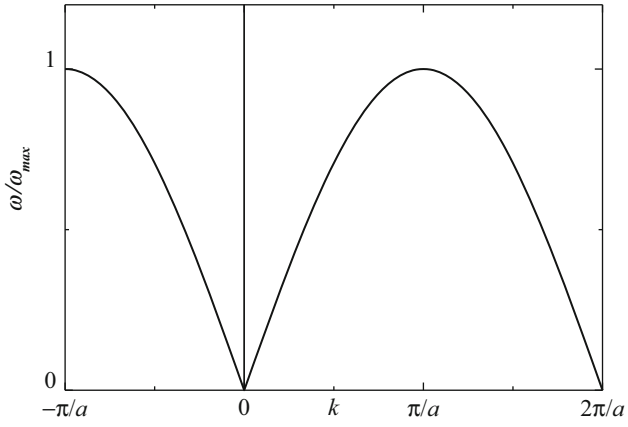


Fig. 6.3 Phonon dispersion relation in a one-dimensional monatomic harmonic crystal, expressed through the dependence of the angular frequency as a function of the wavenumber k

characterizing the lattice vibrations are not specific to one particular atom but are rather a property of the entire lattice. As such, the term phonon is used to designate lattice vibrations, and a frequency and a wavenumber are associated with each phonon. A more detailed discussion on phonons can be found in Sect. 6.1.6.

For a small wavenumber ($k \rightarrow 0$), i.e., in the long wave limit, Eq. (6.14) becomes:

$$\omega = \omega_{\max} \frac{a}{2} k \tag{6.15}$$

where we have used the approximation for the sine function, $\sin(x) \approx x$, for $x \rightarrow 0$, which is in fact the Taylor expansion of the sine function near zero (see Eq. (6.1)). Equation (6.15) means that the angular frequency ω is proportional to the wavenumber k in the long wave limit. Neighboring atoms have similar displacements in this region.

In the short wavelength limit, as k increases, the slope of ω decreases and becomes flat at the zone boundaries $k = \pm\pi/a$. At this point, the atoms in adjacent cells are vibrating with opposite phase. In other words, alternate springs are compressed and stretched, giving rise to maximum atomic displacement and frequencies of vibration.

6.1.4 One-Dimensional Diatomic Harmonic Crystal

Formalism

In the previous sections, we have discussed the motion of atoms in a one-dimensional monatomic crystal where all the atoms are identical, with a mass M, and their equilibrium positions are equally spaced (spacing a). In crystallography terms, we considered a basis of one atom per unit cell. A more general description of atomic motion in a crystal involves a basis with more than one atom.

In this section, we will consider a one-dimensional diatomic harmonic crystal. Ionic crystals such as NaCl and CsCl, atomic crystals such as Si and Ge, and binaries such as GaAs and InP are examples of lattices whose unit cells contain two atoms each. The following parameters need to be introduced for a complete diatomic model. The masses of the two different atoms (labeled 1 and 2) in a unit cell will be denoted M_1 and M_2, respectively, with $M_1 > M_2$. The equilibrium distance between the two atoms in a unit cell is generally arbitrary, but we will choose it to be half the primitive unit cell length for simplicity, i.e., $a/2$. In addition, the elastic force constant C, as defined inEq. (6.2), should be different depending on if an atom interacts with its front or its back neighbor. But for simplicity, we will consider only one force constant C. In spite of these simplifications, the discussion and the results will not lose their generality, even if the mathematical steps will be significantly simpler.

Each diatomic basis will be indexed by an integer n. The displacement of atom 1 from its equilibrium position will be denoted $u_n(t)$, while the displacement of atom 2 will be denoted $v_n(t)$. The atoms are taken to oscillate in the same direction as the

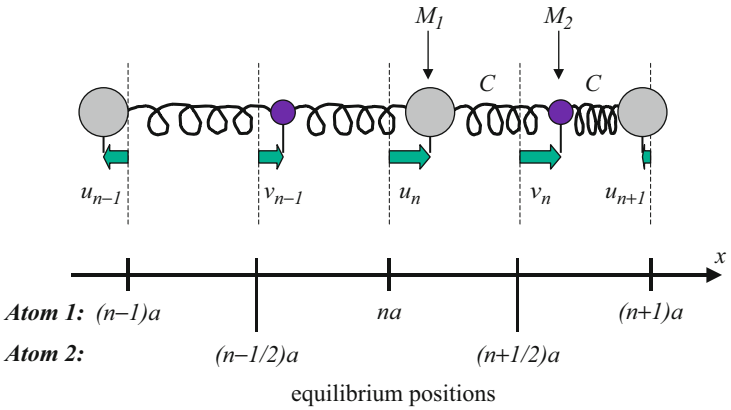


Fig. 6.4 One-dimensional model for the interaction of atoms in a diatomic harmonic crystal structure with atom masses M_1 and M_2. It is assumed here that all the springs have the same constant

lattice (i.e., longitudinal vibration). All these parameters and their simplifications are summarized in Fig. 6.4.

Two coupled sets of equations of motion, similar to Eq. (6.5), need to be considered; one for the displacement of the n^{th} atom 1 and one for the displacement of the n^{th} atom 2:

$$\begin{cases} M_1 \dfrac{d^2 u_n}{dt^2} = -C(2u_n - v_{n-1} - v_n) \\ M_2 \dfrac{d^2 v_n}{dt^2} = -C(2v_n - u_n - u_{n+1}) \end{cases} \tag{6.16}$$

Here again, we seek solutions to the set of Eq. (6.16) in the form of traveling waves with the same wavenumber k and angular frequency ω:

$$\begin{cases} u_n(t) = A\exp[i(kan - \omega t)] \\ v_n(t) = B\exp[i(ka(n + \frac{1}{2}) - \omega t)] \end{cases} \tag{6.17}$$

where A and B are the amplitude of the displacements.

Phonon Dispersion Relation

Substituting these traveling wave expressions into Eq. (6.16), we obtain:

$$\begin{cases} -M_1\omega^2 A\exp[i(kan - \omega t)] \\ = -C(2A\exp[i(kan - \omega t)] - B\exp[i(ka(n - \frac{1}{2}) - \omega t)] - B\exp[i(ka(n + \frac{1}{2}) - \omega t)]) \\ -M_2\omega^2 B\exp[i(ka(n + \frac{1}{2}) - \omega t)] \\ = -C(2B\exp[i(ka(n + \frac{1}{2}) - \omega t)] - A\exp[i(kan - \omega t)] - A\exp[i(ka(n + 1) - \omega t)]) \end{cases}$$

Dividing by $\exp[i(kan - \omega t)]$ the first expression and $\exp[i(ka(n + ½) - \omega t)]$ the second expression, we get:

$$\begin{cases} -M_1\omega^2 A = -C\left(2A - B\exp\left(-\dfrac{ika}{2}\right) - B\exp\left(+\dfrac{ika}{2}\right)\right) \\ -M_2\omega^2 B = -C\left(2B - A\exp\left(-\dfrac{ika}{2}\right) - A\exp\left(+\dfrac{ika}{2}\right)\right) \end{cases}$$

After rearranging the terms with A and those with B:

$$\begin{cases} A[2C - M_1\omega^2] - BC\left[\exp\left(-\dfrac{ika}{2}\right) + \exp\left(+\dfrac{ika}{2}\right)\right] = 0 \\ -AC\left[\exp\left(-\dfrac{ika}{2}\right) + \exp\left(+\dfrac{ika}{2}\right)\right] + B[2C - M_2\omega^2] = 0 \end{cases}$$

Expressing the sum of exponentials with trigonometric functions, we get:

$$\begin{cases} A[2C - M_1\omega^2] - B\left[2C\cos\left(\dfrac{ka}{2}\right)\right] = 0 \\ -A\left[2C\cos\left(\dfrac{ka}{2}\right)\right] + B[2C - M_2\omega^2] = 0 \end{cases} \tag{6.18}$$

This system of equation has a nonzero solution, i.e., A and B not both equal to zero, if and only if the determinant of the system is zero:

$$\left[2C - M_1\omega^2\right]\left[2C - M_2\omega^2\right] - \left[2C\cos\left(\frac{ka}{2}\right)\right]\left[2C\cos\left(\frac{ka}{2}\right)\right] = 0 \tag{6.19}$$

which, after developing the products, becomes:

$$M_1M_2\omega^4 - 2C(M_1 + M_2)\omega^2 + 4C^2 - 4C^2\cos^2\left(\frac{ka}{2}\right) = 0$$

or:

$$M_1M_2\omega^4 - 2C(M_1 + M_2)\omega^2 + 4C^2\sin^2\left(\frac{ka}{2}\right) = 0 \tag{6.20}$$

This equation is of the form $\alpha\omega^4 - 2\beta\omega^2 + \gamma = 0$, with α, β, and $\gamma > 0$, and has two solutions for ω^2, denoted ω_+^2, and ω_-^2 such that:

$$\omega_\pm^2 = \frac{\beta \pm \sqrt{\beta^2 - \alpha\gamma}}{\alpha} \tag{6.21}$$

Therefore, the solutions of Eq. (6.20) are:

$$\omega_{\pm}^2(k) = \frac{C(M_1+M_2) \pm \sqrt{C^2(M_1+M_2)^2 - 4C^2M_1M\sin^2\left(\frac{ka}{2}\right)}}{M_1M_2}$$

which can be simplified into:

$$\omega_{\pm}^2 = C\left(\frac{M_1+M_2}{M_1M_2}\right) \pm C\sqrt{\left(\frac{M_1+M_2}{M_1M_2}\right)^2 - \frac{4\sin^2\left(\frac{ka}{2}\right)}{M_1M_2}}$$

Using the trigonometric identity $\cos(2x) = 1 - 2\sin^2(x)$, this equation becomes:

$$\omega_{\pm}^2(k) = C\left(\frac{M_1+M_2}{M_1M_2}\right)\left[1 \pm \sqrt{1 - \frac{2M_1M_2}{(M_1+M_2)^2}(1-\cos(ka))}\right] \tag{6.22}$$

which constitutes the phonon dispersion relation in the model considered, similar to that obtained in Eq. (6.14). This expression always has a meaning since the argument of the square root is always positive because we have, for any value of masses M_1 and M_2 and value of wavenumber k:

$$0 \leq (1-\cos(ka)) \leq 2$$

and therefore:

$$0 \leq \frac{2M_1M_2}{(M_1+M_2)^2}(1-\cos(ka)) \leq \frac{4M_1M_2}{(M_1+M_2)^2} \leq 1$$

There are thus two possible dispersion relations, denoted $\omega_+(k)$ and $\omega_-(k)$, relating the angular frequency to the wavenumber. Both are plotted in the first Brillouin zone in Fig. 6.5. These plots represent the so-called phonon spectrum of a one-dimensional diatomic harmonic crystal.

The values for $\omega_+(k)$ and $\omega_-(k)$ at $k = 0$ and $k = \pm\frac{\pi}{a}$ can be easily calculated from Eq. (6.22) (note that we have chosen $M_1 > M_2$). The top curve in Fig. 6.5 corresponds to $\omega_+(k)$ and is called the optical phonon branch or simply optical phonon, while the bottom branch corresponds to $\omega_-(k)$ and is called the acoustic phonon branch or simply acoustic phonon.

Now, for small values of wavenumber ($k \to 0$), an approximate expression can be derived from Eq. (6.22). To do so, we start by using an approximate expression for the cosine function in the Eq. (6.22):

$$\cos(ka) \approx 1 - \frac{1}{2}(ka)^2$$

This approximation is in fact the Taylor expansion of the cosine function near zero (see Eq. (6.1)). We therefore obtain successively:

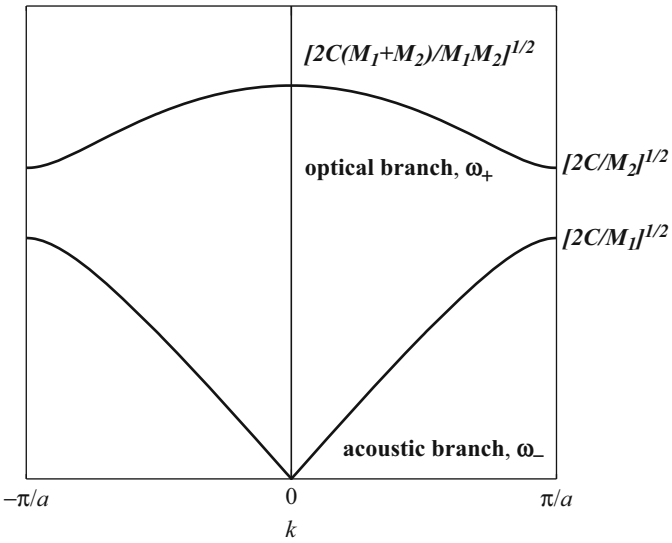


Fig. 6.5 Optical and acoustic branches in the dispersion relation

$$1 - \cos(ka) \approx \frac{1}{2}(ka)^2$$

$$\sqrt{1 - \frac{2M_1M_2}{(M_1 + M_2)^2}(1 - \cos(ka))} \approx \sqrt{1 - \frac{M_1M_2}{(M_1 + M_2)^2}(ka)^2}$$

$$\text{and } \sqrt{1 - \frac{2M_1M_2}{(M_1 + M_2)^2}(1 - \cos(ka))} \approx 1 - \frac{M_1M_2}{2(M_1 + M_2)^2}(ka)^2$$

by using the approximation $\sqrt{1-x} \approx 1 - \frac{1}{2}x$ for $x \to 0$ (again this comes from the Taylor expansion of $\sqrt{1-x}$ for small values of x). Equation (6.22) can then be approximated by the following expression:

$$\omega_{\pm}^2(k) \approx C\left(\frac{M_1 + M_2}{M_1M_2}\right)\left[1 \pm \left(1 - \frac{M_1M_2}{2(M_1 + M_2)^2}(ka)^2\right)\right] \tag{6.23}$$

Consequently, in the long wave limit, the angular frequency of the acoustic phonon branch can be written as:

$$\omega_{-}^2(k) \approx C\left(\frac{M_1 + M_2}{M_1M_2}\right)\left[\frac{M_1M_2}{2(M_1 + M_2)^2}(ka)^2\right]$$

$$\omega_{-}(k) \approx k\sqrt{\frac{Ca^2}{2(M_1+M_2)}} \tag{6.24}$$

which means that the angular frequency $\omega_{-}(k)$ in the acoustic phonon branch is proportional to the wavenumber k, similar to the result obtained in Eq. (6.15). The shape of the acoustic branch is similar, but the increased mass lowers the frequency. For the acoustic branch in the long wave limit, the traveling wave is equivalent to the elastic wave of a one-dimensional atomic chain regarded as a continuous media. The nature of the vibrations in this region is just like sound waves. The two atoms in the unit cell move in the same direction, and over a small region, it seems as if the entire crystal has been compresses or stretched. This is why the $\omega_{-}(k)$ branch is called the acoustic branch.

In the same limit ($k \to 0$), the angular frequency of the optical phonon branch can be expressed from Eq. (6.23):

$$\omega_{+}^2(k) \approx C\left(\frac{M_1+M_2}{M_1M_2}\right)[1+1] = 2C\left(\frac{M_1+M_2}{M_1M_2}\right) \tag{6.25}$$

which shows that the angular frequency $\omega_{+}(k)$ in the optical phonon branch is constant in the long wave limit. The nature of the vibrations in this region is that the two atoms in the unit cell move in opposite directions. This is similar to the top of the band in the monatomic case, where there is maximum distortion and frequency of vibration. The angular frequency in the limit ($k \to \pi/a$) for the optical and acoustic branches is left as an exercise at the end of the chapter.

Furthermore, the ratio of the displacement amplitudes A and B defined in Eq. (6.17) can be taken for two different values, depending on the branch chosen, calculated from either one of Eq. (6.18):

$$\left(\frac{B}{A}\right)_{\pm} = \frac{2C - M_1\omega_{\pm}^2}{2C\cos\left(\frac{ka}{2}\right)} \tag{6.26}$$

Again, in the long wave limit ($k \to 0$) and for the acoustic phonon branch, we have $\omega_{-}(k) \to 0$ as seen from Eq. (6.24) and $\cos\left(\frac{ka}{2}\right) \to 1$ so that:

$$\left(\frac{B}{A}\right)_{-} \to \frac{2C}{2C} = 1 \tag{6.27}$$

which demonstrates that, in this case, the vibrations of the two atoms in one primitive unit cell have exactly the same amplitude and phase (i.e., direction), as shown in Fig. 6.6.

In the long wave limit ($k \to 0$) for the optical phonon branch, we have $\omega_{+} \to \sqrt{\frac{2C}{\left(\frac{M_1M_2}{M_1+M_2}\right)}}$ from Eq. (6.25), and therefore, by substituting into Eq. (6.):

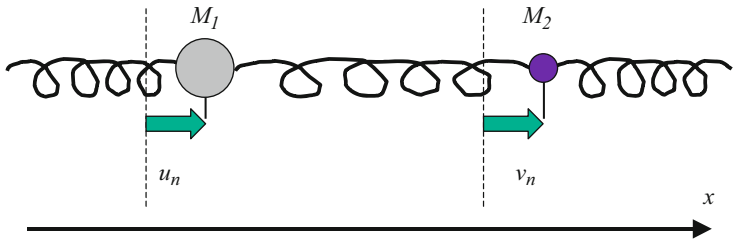


Fig. 6.6 Atomic vibrations in a one-dimensional diatomic harmonic crystal, corresponding to the acoustic phonon branch. In this configuration, the two atoms forming the unit cell move in the same direction at the same time

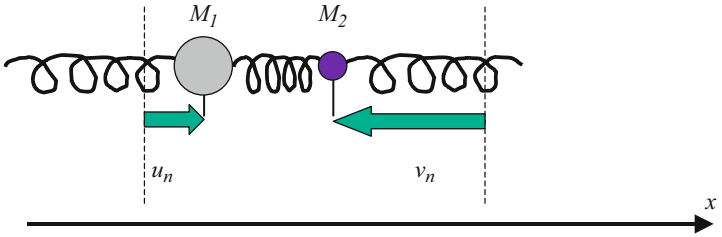


Fig. 6.7 Atomic vibrations in a one-dimensional diatomic harmonic crystal, corresponding to the optical phonon branch. In this configuration, the two atoms forming the unit cell move in opposite directions at the same time

$$\left(\frac{B}{A}\right)_+ \rightarrow \frac{2C - M_1 \frac{2C}{\left(\frac{M_1 M_2}{M_1+M_2}\right)}}{2C} = -\frac{M_1}{M_2} \tag{6.28}$$

which shows that, in the long wave limit of the optical branch, the vibrations of the two atoms in one primitive unit cell have a specific amplitude ratio and opposite phases (i.e., directions), as shown in Fig. 6.7. Thus, optical phonons are described by the oscillations of two atoms about a center of mass, while acoustic phonons are described by the movement of the two atoms center of mass. The amplitude ratio in the limit ($k \rightarrow \pi/a$) is left as an exercise at the end of the chapter.

Actually, the ratio of the amplitudes is such that the vibrations of the two atoms in a primitive unit cell leave the position of their center of gravity unchanged. Therefore, if the two atoms are ions of opposite charges, such as in the case of GaAs or NaCl, these oscillations result in a periodic oscillation of the amplitude of the dipole moment formed by these two charged ions, as discussed in Sect. 1.5.6. Such oscillations of the dipole moment are frequently optically active, i.e., are involved in the absorption or emission of electromagnetic (infrared mostly) radiation. This explains the use of the term "optical" for the $\omega_+(k)$ branch of lattice vibrations.

One can use the dispersion relation for phonons and photons to examine the conservation of energy and momentum that applies to the interaction of phonons and

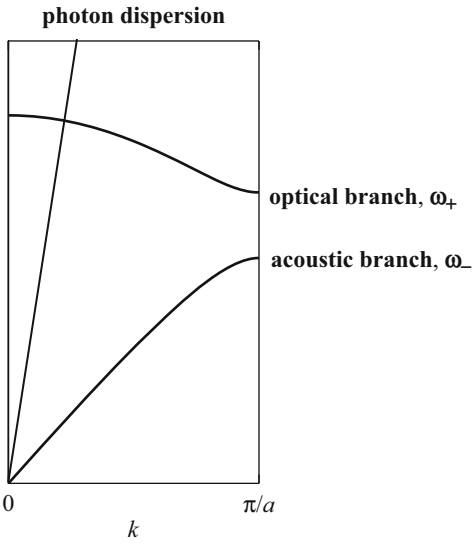


Fig. 6.8 The dispersion curves for a photon and an acoustic and optical phonon. The optical branch crosses with the photon branch, allowing for energy and momentum conservation

photons. Figure 6.8 shows the crossing of the dispersion relation for both acoustic and optical phonons with a photon. Because the photon and optical phonon curves cross, energy and momentum can be exchanged. An optical phonon can be created or annihilated with a photon. Since the acoustic mode never crosses the photon dispersion, they cannot interact. For example, in NaCl, its optical mode is excited by light because an electric field can displace the two oppositely charged ions in different directions. In a Ge crystal, the two atoms in the unit cell have similar charges and cannot be excited by an electric field.

6.1.5 Extension to Three-Dimensional Case

Formalism

So far, we have only considered a one-dimensional atomic crystal. A real crystal expands in all three dimensions of space, and lattice vibrations are more complicated. For example, the vibrations can occur in all three directions, regardless of the equilibrium position alignment of the atoms, and need to be expressed using a displacement vector $\vec{u}_{\vec{R}}(t)$. Moreover, a wavevector $\vec{k}$ must be used, similarly to the way it was done in Chap. 5 for three-dimensional electronic band structures. This wavevector $\vec{k}$ also indicates the direction of propagation of the traveling wave. The expression of the displacement, given for the one-dimensional case in Eq. (6.6), becomes in the three-dimensional case now:

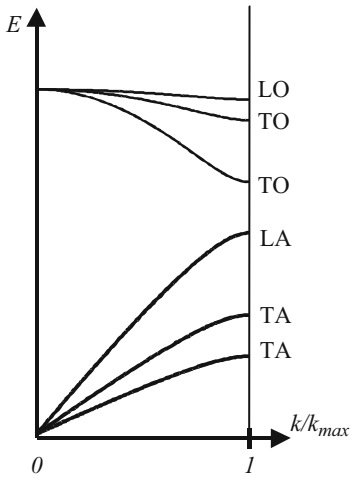


Fig. 6.9 Typical phonon dispersion spectrum for a three-dimensional diatomic lattice ($s = 2$)

$$\vec{u}_{\vec{R}}(t) = \vec{A} \exp\left[i\left(\vec{k} \cdot \vec{R} - \omega t\right)\right] \tag{6.29}$$

where $\vec{A}$ is the amplitude vector of the displacement and $\vec{k} \cdot \vec{R}$ is the dot product between the wavevector and the equilibrium position $\vec{R}$ of the atom considered.

In spite of this increased complexity, all the features obtained in the present simplified study remain valid. In particular, there still exist two types of phonons, as shown in the example of dispersion spectrum in Fig. 6.9: acoustic phonons, for which the vibration frequency goes to zero in the long wave limit ($\left|\vec{k}\right| \to 0$), and optical phonons, for which the frequency goes to a nonzero finite value in the long wave limit. Each type of phonons is further divided into two main categories: transversal and longitudinal phonons. The terms "transversal" and "longitudinal" refer to the direction of atomic displacements $\vec{u}(t)$ with respect to direction of propagation $\vec{k}$: perpendicular for transversal and parallel for longitudinal. There are generally two transverse and one longitudinal branch for each optical and acoustic phonons. Furthermore, the dispersion relations are not always isotropic, meaning that the phonon dispersion relations are different for different symmetry directions within the crystal.

For example, in Fig. 6.9, the transversal acoustic (TA), longitudinal acoustic (LA), transversal optical (TO), and longitudinal optical (LO) phonon branches are shown. Notice that the longitudinal branches are higher in energy than the transverse branches. In general, for a three-dimensional crystal with s atoms per unit cell, there are always three acoustic branches, two transversal and one longitudinal. There are also $3s-3$ optical branches. Figure 6.9 shows a typical example for $s = 2$. A monatomic Bravais lattice ($s = 1$) can only have acoustic phonon branches.

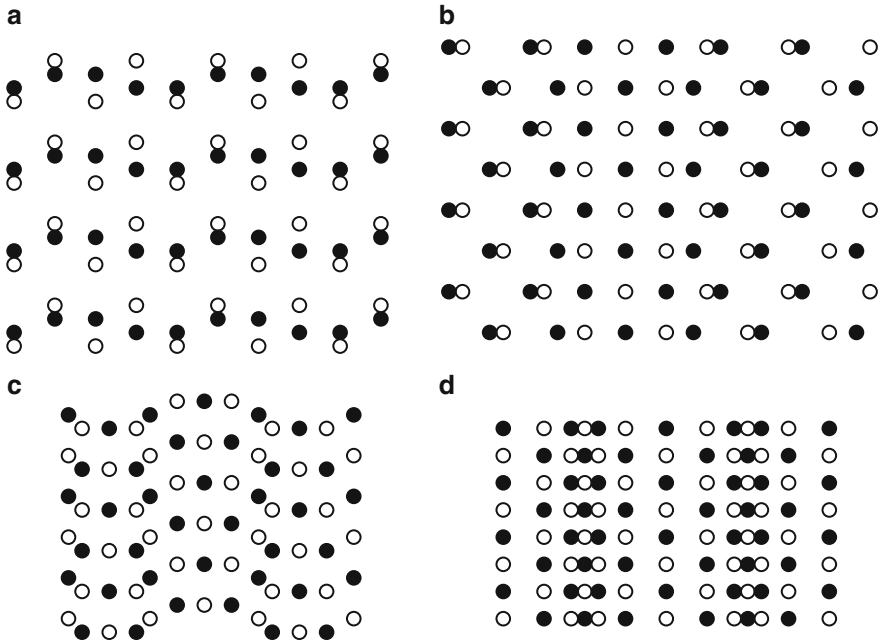


Fig. 6.10 The propagation of the four different phonon modes through a lattice: (**a**) transverse optic, (**b**) longitudinal optic, (**c**) transverse acoustic, and (**d**) longitudinal acoustic

Figure 6.10 shows the movement of (a) transverse optic (TO), (b) longitudinal optic (LO), (c) transverse acoustic (TA), and (d) longitudinal acoustic (LA) phonons in a lattice. The black circles represent the atoms with smaller mass, such as the gallium atoms in gallium arsenide. The white circles represent the heavier atoms, such as the arsenic atoms in gallium arsenide.

TO phonons propagate by the lighter atoms (black) being displaced perpendicular to the direction of the wave traveling. The heavier atoms (white) remain somewhat stationary within the lattice. For LO phonons, the heavier atoms remain somewhat stationary within the lattice, while the lighter atoms move parallel to the propagation of the traveling wave. As you can see, both optic modes produce a change in dipole movement, or the movement of the atoms about their center of mass. The heavier atoms remain fixed in the lattice, while the lighter atoms move and carry the wave through the medium. TA modes propagate similar to a pulse moving along a string after it has been jerked. The wave propagates through the movement of both the heavier and lighter atoms. Lastly, LA phonons propagate through the movement of a pair of atoms toward and away from another pair of atoms. Both acoustic modes correspond to the movement of the center of mass of two atoms. The distance between a heavier and lighter atoms remains fixed, while the pair as a whole is displaced relative to other atom pairs.

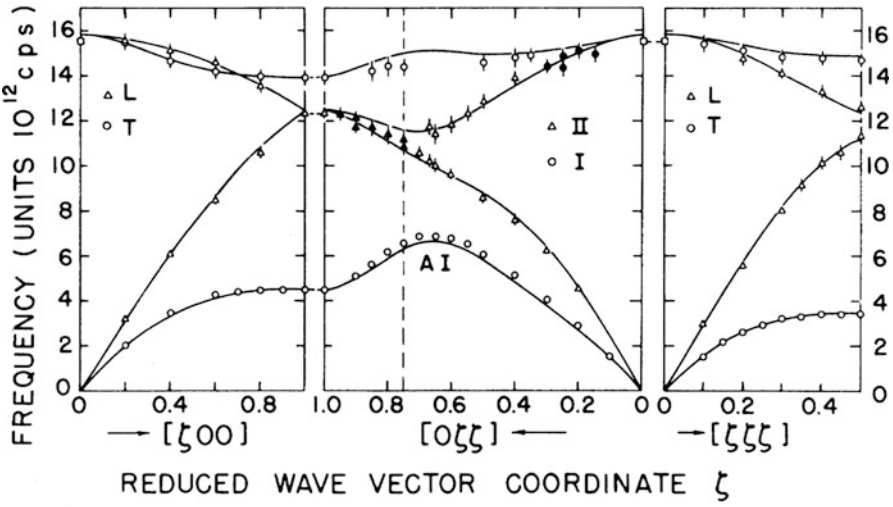


Fig. 6.11 Phonon dispersion relation for silicon in three crystal directions. Solid lines are calculated. Data points: open circles represent transverse (T) modes, open triangles longitudinal (L) modes, and solid points undetermined polarization modes (Reprinted with permission from Dolling 1963, Fig. 1. Copyright 1963, International Atomic Energy Association)

Silicon

Silicon crystals only have two identical atoms in their unit cell and bonds in the diamond structure. This results in the LO and TO energies being degenerate at the zone center. Since both atoms are identical, the bonds do not carry any electronegativity, and there is no restoring force like that in GaAs (Fig. 6.11).

Gallium Arsenide

In GaAs, the LO phonons have higher energy than the TO phonons near the zone center. This results from the ionic nature of the bonding in zinc-blende crystals. In GaAs, the arsenic atoms contribute five electrons to the bonds compared to gallium atoms, which contribute three. Consequently, the electrons spend on average more time near the arsenic atoms resulting the arsenic atoms to be slightly more negative, while the gallium atoms are slightly positive. This difference in electronegativity produces a restoring force for a propagating LO mode but not a TO mode. This increase in energy gives the LO modes a higher frequency (Fig. 6.12).

6.1.6 Phonons

In Chap. 5, the treatment of the electrons in a crystal led to energy levels and momenta that do not correspond to those of individual atoms but are properties of the lattice as a whole. Earlier in this chapter, we have hinted that the characteristics of the traveling waves arising from lattice vibrations are not specific to one particular

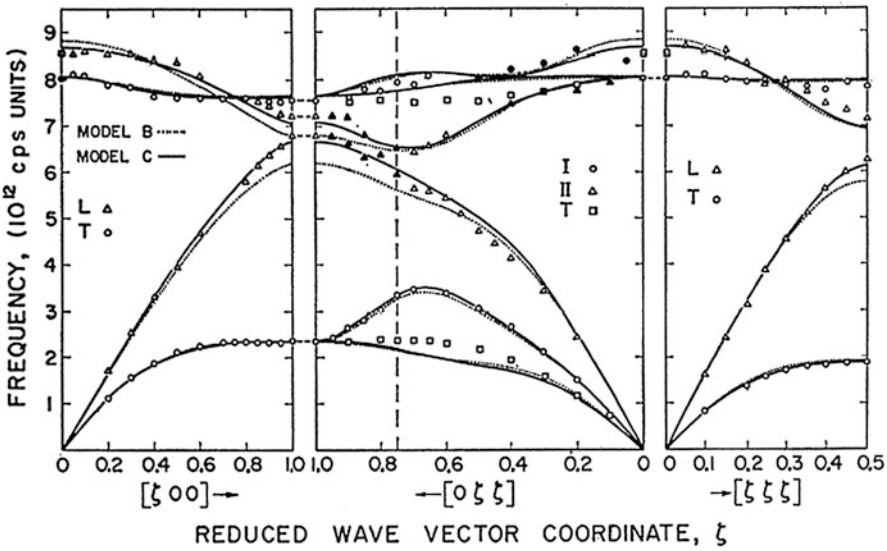


Fig. 6.12 Phonon dispersion relation for gallium arsenide in three crystal directions. Dotted and solid lines denote calculated values. Solid points denote undetermined polarization modes (Reprinted figure with permission from Waugh and Dolling 1963, Fig. 1. Copyright 1963 by the American Physical Society)

atom but are rather a property of the entire lattice too. We thus have to consider the collective excitation of the crystal as a whole and talk about a lattice wave. Each type of vibration is called a vibrational mode and is characterized by a wavevector $\vec{k}$ and a frequency $\omega(\vec{k})$.

The previous sections of this chapter dealt with a classical analysis of lattice vibrations. In a quantum mechanical treatment, especially when lattice waves interact with other objects (e.g., electrons, electromagnetic waves, or photons), it is convenient to regard a lattice wave as a quasiparticle or phonon with a momentum and a (quantized) energy such that:

$$\begin{cases} \vec{p} = \hbar \vec{k} \\ E = \hbar\omega(\vec{k}) \end{cases} \tag{6.30}$$

This is analogous to the quantization of the electromagnetic field discussed in Chap. 4. The energy in Eq. (6.30) is the quantum unit of vibrational energy at that frequency. Because phonons involve vibrational energy stored in the crystal, phonons can interact with other waves or particles such as electrons, photons, and phonons. These types of interactions lead to the experimentally observable physical properties of crystals.

The velocity of a phonon is given by the group velocity of the corresponding traveling wave, defined as the gradient of the frequency with respect to the wavevector:

$$\vec{v}_g = \frac{\partial \omega(\vec{k})}{\partial \vec{k}} = \nabla_{\vec{k}} \omega(\vec{k}) \tag{6.31}$$

In Cartesian coordinates with unit vectors $(\vec{x}, \vec{y}, \vec{z})$, this relation can be written as:

$$\vec{v}_g = \frac{\partial \omega(k_x, k_y, k_z)}{\partial k_x} \vec{x} + \frac{\partial \omega(k_x, k_y, k_z)}{\partial k_y} \vec{y} + \frac{\partial \omega(k_x, k_y, k_z)}{\partial k_z} \vec{z} \tag{6.32}$$

In this quantum picture, the propagation of harmonic lattice waves, i.e., up to the second-order term in Eq. (6.2), is equivalent to the free movement of non-interacting phonon quasiparticles, also called "phonon gas," and their description is similar to that of photons.

In particular, any number of identical phonons may be present simultaneously in the lattice, in any of the phonon mode characterized by a wavevector $\vec{k}$ for a given temperature. A phonon gas thus obeys the Bose-Einstein statistics which says that the average number of phonons in a given mode $(\vec{k})$ is then determined by:

$$N_{\vec{k}} = \frac{1}{\exp\left(\frac{\hbar\omega(\vec{k})}{k_b T}\right) - 1} \tag{6.33}$$

where k_b is the Boltzmann constant and T is the absolute temperature. At high temperatures, i.e., $k_b T >> \hbar\omega(\vec{k})$, the exponential in Eq. (6.33) can be approximated by:

$$\exp\left(\frac{\hbar\omega(\vec{k})}{k_b T}\right) \approx 1 + \frac{\hbar\omega(\vec{k})}{k_b T} \tag{6.34}$$

where we have used the approximation $\exp(x) \approx 1 + x$ for $x \to 0$ (again this comes from the Taylor expansion of $\exp(x) \approx 1 + x$ for small values of x). Therefore, $N_{\vec{k}} \approx \frac{k_b T}{\hbar\omega(\vec{k})}$, which expresses that the average number of phonons in a given mode is proportional to the temperature, at high temperatures.

As mentioned earlier, phonons can interact with other phonons. Such interaction would correspond to anharmonic vibrations in the classical wave picture, which arise from cubic and higher order terms in Eqs. (6.1 and 6.2).

Example

Q Estimate the average number of phonons in a given mode at low temperatures.

A The average number of phonons $N(E)$ with an energy E is given by: $N(E) = \frac{1}{\exp\left(\frac{E}{k_b T}\right) - 1}$. At low temperatures, we have $\exp\left(\frac{E}{k_b T}\right) >> 1$, and the expression for $N(E)$ can be simplified into: $N(E) \approx \exp\left(-\frac{E}{k_b T}\right)$.

6.1.7 Sound Velocity

It is known that a solid can transmit sound. This is in fact accomplished through the vibrations of atoms similar to the ones discussed in earlier sections. The sound velocity is the speed at which sound propagates and is related to velocity of a traveling wave as discussed below.

In Sect. 6.1.3, we have already hinted that the velocity of the traveling wave was given by the ratio of the angular frequency to the wavenumber in Eq. (6.8):

$$v_{\text{ph}} = \frac{\omega}{k} \tag{6.35}$$

Using Eqs. (6.13 and 6.14), we obtain:

$$v_{\text{ph}} = \sqrt{\frac{4C}{M}} \left| \frac{\sin\left({}^{ka}/_{2}\right)}{k} \right| = a\sqrt{\frac{C}{M}} \left| \frac{\sin\left({}^{ka}/_{2}\right)}{{}^{ka}/_{2}} \right| = v_0 \left| \frac{\sin\left({}^{ka}/_{2}\right)}{{}^{ka}/_{2}} \right| \tag{6.36}$$

where:

$$v_0 = a\sqrt{\frac{C}{M}} \tag{6.37}$$

Therefore:

$$v_{\text{ph}} = v_0 \left| \frac{\sin\left({}^{ka}/_{2}\right)}{{}^{ka}/_{2}} \right| \tag{6.38}$$

This quantity is called the phase velocity because it represents the velocity of the phase of the wave or, in other words, the speed at which the peak of the wave travels in space. The phase velocity is plotted in Fig. 6.13, and we see that it never reaches zero.

There is another quantity of interest which is the group velocity of a traveling wave which represents the velocity of a wave packet and therefore of the wave energy and is defined as:

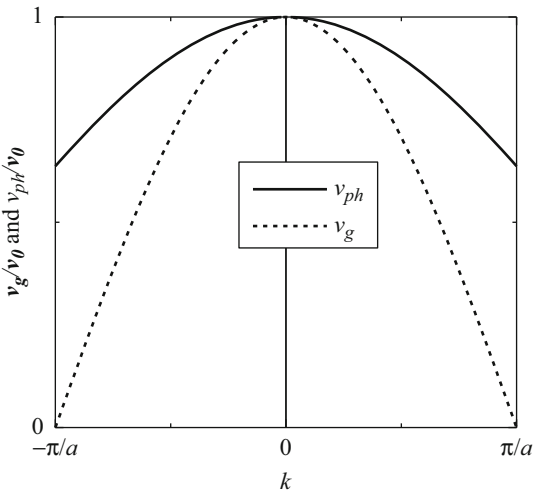


Fig. 6.13 Phase and group velocities versus wavenumber k

$$v_g = \left|\frac{d\omega}{dk}\right| \tag{6.39}$$

Using Eqs. (6.13 and 6.14), we obtain:

$$v_g = \sqrt{\frac{4C}{M}}\frac{a}{2}\left|\cos\left(\tfrac{ka}{2}\right)\right| = a\sqrt{\frac{C}{M}}\left|\cos\left(\tfrac{ka}{2}\right)\right| \tag{6.40}$$

and therefore:

$$v_g = v_0\left|\cos\left(\tfrac{akl}{2}\right)\right| \tag{6.41}$$

The group velocity is also plotted in Fig. 6.6. We see that this quantity drops to zero when $k \to \frac{\pi}{a}$, i.e., at boundary of the first Brillouin zone.

Example

Q Estimate the order of magnitude for the elastic constant C of silicon, given that the sound velocity in silicon is 2.2×10^5 cm·s^{-1}.

A Starting from the expression for the sound velocity, $v_0 = a\sqrt{\frac{C}{M}}$, where $a = 5.43$ Å and $M = 28M_p$ are the lattice constant and mass of a silicon atom, respectively. We thus have:

$$C = M\frac{v_0^2}{a^2} = \left(28 \times 1.67264 \times 10^{-27}\right) \times \frac{\left(2.2 \times 10^3\right)^2}{\left(5.43 \times 10^{-10}\right)^2}$$
$$\approx 0.77 N.m^{-1}$$

From Eq. (6.37), we see that the speed of sound in a medium is proportional to the inverse square root of M, the atomic mass, and the square root of C, the elastic constant of the material. A generalized form for the speed of sound in a medium is:

$$v_s = \sqrt{\frac{B}{\rho}} \tag{6.42}$$

where B is the bulk modulus of the material and ρ is the density, given by its mass divided by its volume.

The bulk modulus is the property that determines the extent to which a medium changes its volume in response to an applied pressure. A generalized expression for the bulk modulus of a material is given by:

$$B = -\frac{\Delta p}{\frac{\Delta V}{V}} \tag{6.43}$$

where p is an applied pressure and V is the medium's volume. $\Delta V/V$ is the percent change in volume produced by a change in pressure Δp. The minus sign is included because whenever we increase the pressure, the volume decreases and vice versa. The minus sign allows what is under the radical in Eq. (6.42) to be positive.

Just as phonon modes can be anisotropic in a crystal, the bulk modulus is also directional within a crystal, and the velocity of sound is dependent upon what direction the sound is traveling in a material. A medium's bulk modulus generally takes on a tensor form and can be significantly different in the Γ, X, and L directions. This results from the crystal structure (e.g., cubic, tetragonal, orthorhombic, etc.) having different bonding lengths on different sides of each atom.

6.1.8 Summary

In this chapter, we have described the basic formalism for treating the interaction between atoms in a crystal, through the simple examples of one-dimensional monatomic and diatomic harmonic lattices. Several important concepts have been introduced such as the lattice vibrational modes, traveling waves, dispersion relations, acoustic and optical branches, longitudinal and transversal branches, and sound velocity. We realized that these lattice vibrations could be quantized in the same manner as the electromagnetic field and can thus be considered as quasiparticles, or phonons, with a momentum and energy and which obey Bose-Einstein statistics.

6.2 Thermal Properties of Crystals

6.2.1 Introduction

In Chap. 6 Part 1, we built simple mathematical models to describe the vibrations of atoms, first in a one-dimensional system and then extended to a three-dimensional harmonic crystal. These models, in the quantum description, led us to introduce a quasiparticle called the phonon, with an associated momentum and energy spectrum. Many of the phenomena measured in crystals can be traced back to phonons.

In this chapter, we will employ the results of the phonon formalism used in Chap. 6 to interpret the thermal properties of crystals, in particular their heat capacity, thermal expansion, and thermal conductivity.

6.2.2 Phonon Density of States (Debye Model)

Debye Model

The Debye model was developed in the early stages of the quantum theory of lattice vibration in an effort to describe the observed heat capacity of solids (Sect. 6.1.3). The model relies on a simplification of the phonon dispersion relation (see, e.g., Eq. (6.22), Fig. 6.5, or Fig. 6.8). In the Debye model, all the phonon branches are replaced with three acoustic branches, one longitudinal (l) and two transversal (t), with corresponding phonon spectra:

$$\omega_n\left(\vec{k}\right) = v_n\left|\vec{k}\right| = v_n k \tag{6.44}$$

where n ($= l$ or t) is an index, k is the norm or length of the wavevector $\vec{k}$, and v_l and v_t are the longitudinal and transversal sound velocities, respectively. This model corresponds to a linearization of the phonon spectrum as shown in Fig. 6.4. But this linearization implies that the phonon frequencies depend solely on the norm of the wavevector. Some boundary conditions therefore need to be changed in this model (Fig. 6.14).

Indeed, we remember that the range for the wavevector was restricted to the first Brillouin zone in the real phonon dispersion relation. The Born-von Karman boundary conditions of Sect. 5.3 limited the total number of allowed values for $\vec{k}$ to the number N of atoms in the crystal of volume V considered. We saw in Sect. 5.3 that the volume occupied by each wavevector was $\frac{(2\pi)^3}{V}$. The volume of the first Brillouin zone is then $\frac{2(\pi)^3 N}{V}$ and must be equal to $\frac{4\pi}{3}k_D^3$ where k_D is the Debye wavenumber such that the relation (7.1) is valid in the range $0 \leq k \leq k_D$. We thus obtain:

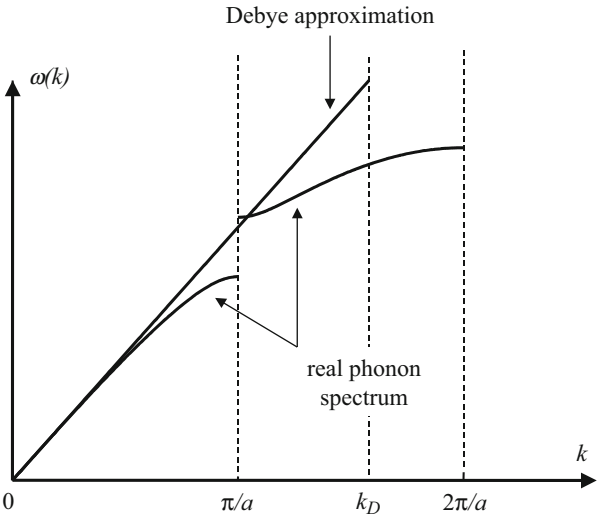


Fig. 6.14 Illustration of the Debye model in the phonon dispersion curve. In the Debye model, all the phonon branches are replaced with three acoustic branches. This corresponds to a simplification of the phonon dispersion spectrum, through a linearization of the phonon branches. A sphere is defined in momentum space with radius k_D, the Debye wavevector, such that the total number of modes inside the Debye sphere now matches the total number of modes in the real system

$$k_D^3 = \frac{6\pi^2 N}{V} \tag{6.45}$$

This wavenumber corresponds to a Debye frequency ω_D defined by:

$$\hbar\omega_D = \hbar\nu_0 k_D \tag{6.46}$$

where ν_0 is the sound velocity in the material. The Debye frequency is characteristic of a particular solid material and is approximately equal to the maximum frequency of lattice vibrations. It is also useful to define the Debye temperature Θ_D such that:

$$k_b\Theta_D = \hbar\omega_D = \hbar\nu_0 k_D \tag{6.47}$$

The significance of Θ_D will become clear in the following discussion. However, it follows that every solid will have its own characteristic phonon spectrum and therefore its own Debye temperature. The Debye temperatures for a few solids are listed in Table 6.1.

Example

Q Calculate the Debye wavelength for GaAs, given that the density of GaAs is $d = 5.32 \times 10^3$ kg·m^{-3}.

A We make use of the expression giving the Debye wavenumber $k_D^3 = \frac{6\pi^2 N}{V}$, which is related to the Debye wavelength through $\lambda_D = \frac{2\pi}{k_D} = 2\pi\left(\frac{6\pi^2 N}{V}\right)^{-1/3}$, where N is the number of atoms in the volume V. By definition of the density,

Table 6.1 Debye temperatures of a few solids (Grigoriev and Meilikhov 1997)

Material	Θ_D (K)
Pb	105
Au	162
Ag	227
NaCl	275
GaAs	345
Cu	347
Ge	373
W	383
Al	433
Fe	477
Si	650
BN	1900
C (diamond)	2250

we have $d = \frac{1}{V}\frac{N}{2}(M_{Ga} + M_{As})$, where M_{Ga} and M_{As} are the masses of a Ga and an As atom, respectively. The factor 2 arises from the fact that half of the atoms in the volume are Ga atoms and the other half are As atoms.

Therefore, we can write:

$$\lambda_D = 2\pi\left(6\pi^2 \frac{2d}{(M_{Ga}+M_{As})}\right)^{-1/3}$$
$$= 2\pi\left(6\pi^2 \frac{2\times 5.32\times 10^3}{(69.7+74.9)\times 1.67264\times 10^{-27}}\right)^{-1/3}$$

or $\lambda_D = 4.57\,\text{Å}$.

Phonon Density of States

The phonon density of states $g(\omega)$ is the number of phonon modes $\vec{k}$ per unit frequency interval which have a frequency $\omega(\vec{k})$ equal to a given value ω. It can be calculated in a way similar to that used for the electron density of states in Sect. 6.1.3:

$$g(\omega) = \sum_{\vec{k},n} \delta\left[\omega_n(\vec{k}) - \omega\right] \tag{6.48}$$

where the summation is performed over all phonon modes $\vec{k}$ and phonon branches labeled n. Because the crystal has macroscopic sizes, the strictly discrete wavevector $\vec{k}$ can be considered quasi-continuous, as was done in Chap. 6 Eq. 6.44, and the discrete summation can be replaced by an integral:

$$\sum_{\vec{k}} \Upsilon\left(\vec{k}\right) \equiv \frac{V}{(2\pi)^3} \int_{k} \Upsilon\left(\vec{k}\right) d\,\vec{k} = \frac{V}{(2\pi)^3} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Upsilon\left(k_x, k_y, k_z\right) dk_x dk_y dk_z \tag{6.49}$$

where V is the volume of the crystal considered. The summation here is actually performed over all values of $\vec{k}$ in the first Brillouin zone. Equation (6.48) then becomes:

$$g(\omega) = \frac{V}{(2\pi)^3} \sum_{n} \int\int_{k} \int \delta\left[\omega_n\left(\vec{k}\right) - \omega\right] d\,\vec{k} \tag{6.50}$$

We now make use of Eq. (5.37):

$$d\,\vec{k} = d\left(\frac{4\pi}{3}k^3\right) = 4\pi k^2 dk$$

where k is the norm or length of the wavevector $\vec{k}$. Therefore, Eq. (6.50) becomes:

$$g(\omega) = \frac{4\pi V}{(2\pi)^3} \sum_{n} \int_{0}^{k_D} \delta\left[\omega_n\left(\vec{k}\right) - \omega\right] k^2 dk \tag{6.51}$$

where the integration is now from 0 to the Debye wavenumber k_D, in agreement with the Debye model described earlier. Substituting (6.46), we get successively:

$$g(\omega) = \frac{V}{2\pi^2} \sum_{n} \int_{0}^{k_D} \delta[\nu_n k - \omega] k^2 dk \tag{6.52}$$

or:

$$g(\omega) = \frac{V}{2\pi^2} \sum_{n} \int_{0}^{k_D} \delta[x - \omega] \frac{x^2}{v_n^3} dx \tag{6.53}$$

after the change of variable $x = v_n k$ (and thus $dx = v_n dk$). There is a nonzero solution only if there is a wavenumber k between *0* and k_D such that $x = v_n k = \omega$, and:

$$\begin{cases} g(\omega) = \dfrac{V}{2\pi^2} \displaystyle\sum_{n} \frac{\omega^2}{v_n^3} & for\, 0 \le \omega \le \omega_D \\ g(\omega) = 0 & for\, \omega_D \le \omega \end{cases} \tag{6.54}$$

Remembering that the Debye model takes into account one longitudinal (l) and two transversal (t) modes, we obtain:

$$\begin{cases} g(\omega) = \dfrac{V}{2\pi^2}\left(\dfrac{\omega^2}{v_l^3} + 2\dfrac{\omega^2}{v_t^3}\right) & for\, 0 \le \omega \le \omega_D \\ g(\omega) = 0 & for\, \omega_D \le \omega \end{cases} \tag{6.55}$$

which can also be rewritten as:

$$\begin{cases} g(\omega) = \dfrac{3V\omega^2}{2\pi^2 v_0^3} & for\, 0 \le \omega \le \omega_D \\ g(\omega) = 0 & for\, \omega_D \le \omega \end{cases} \tag{6.56}$$

where:

$$\frac{1}{v_0^3} = \frac{1}{3}\left(\frac{1}{v_l^3} + \frac{2}{v_t^3}\right) \tag{6.57}$$

is the inverse average sound velocity. This phonon density of states is illustrated in Fig. 6.15 where we have a parabolic relation. Although the Debye model is a simple approximation, the choice of k_D ensures that the area under the curve of $g(\omega)$ is the same as for the real curve for the density of states. Moreover, this expression is precise enough to determine the lattice contribution to the heat capacity both at high and low temperatures.

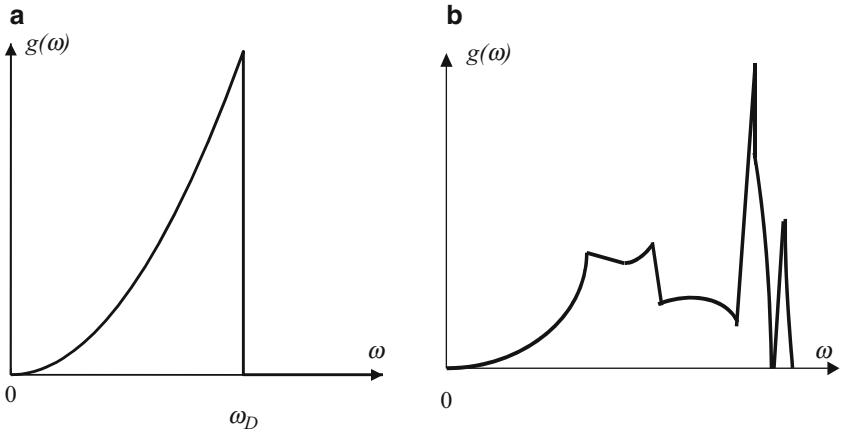


Fig. 6.15 (**a**) Illustration of the phonon density of states in the Debye model, where the relationship is parabolic until the Debye frequency is reached, after which the density of states is equal to zero. (**b**) Illustration of a typical phonon spectrum of a real crystal with discontinuities due to singularities in the spectrum. The singularities are due to zeroes in the group velocity

Heat Capacity

Lattice Contribution to the Heat Capacity (Debye Model)

When heat is transferred to a solid, its temperature increases. Heat has a mechanical equivalent which is an energy and is generally expressed in units of calorie with 1 calorie corresponding to 4.184 joules. Different substances need different amounts of heat energy to raise their temperature by a set amount. For example, it takes 1 calorie to raise 1 g of water by 1 degree K. The same amount of energy, however, raises 1 g of copper by about 11 K.

The heat capacity, C, of a material is a measure of the ability with which a substance can store this heat energy and is described by the ratio of the energy dE transferred to a substance to raise its temperature by an amount dT. The greater a given material's heat capacity, the more energy must be added to change its temperature. The heat capacity is characteristic of a given substance, and its units are cal·K^{-1} or J·K^{-1}. The heat capacity is defined as:

$$C_v = \left(\frac{dE}{dT}\right)_v \tag{6.58}$$

subscripts denoting which variable (volume or pressure) is held constant.

The specific heat capacity, often known simply as the specific heat and denoted by a lowercase c, of a material is the heat capacity per unit the mass. The specific heat of a given substance has units or cal·g^{-1}·K^{-1} or J·kg^{-1}·K^{-1} and is thus specific to a particular material and independent of the quantity of material. A few values of specific heat for elements in the periodic table are given in Fig. A. in Appendix A.3.

Both heat capacity and specific heat phenomena are closely related to phonons because, when a solid is heated, the atomic vibrations become more intense and more phonons or vibrational modes are accessible. A measure of the heat energy received by a solid is therefore the change in the total energy carried by the lattice vibrations. This total energy E can be easily expressed using the following integral, knowing the average number of phonons $N(\omega)$) (Eq. 6.33), the phonon density of states $g(\omega)$, and that a phonon with frequency ω has an energy $\hbar\omega$ (Eq. 6.30):

$$E = \int_0^\infty N(\omega) g(\omega) \hbar\omega d\omega \tag{6.59}$$

In the Debye model, we can use Eq. (6.56) for $g(\omega)$ and rewrite Eq. (6.59) as:

$$E = \int_0^{\omega_D} \frac{1}{\exp\left(\frac{\hbar\omega}{k_b T}\right) - 1} \frac{3V\omega^2}{2\pi^2 v_0^3} \hbar\omega d\omega$$

or:

$$E = \frac{3V\hbar}{2\pi^2 v_0^3} \int_0^{\omega_D} \frac{\omega^3}{\exp\left(\frac{\hbar\omega}{k_b T}\right) - 1} d\omega \tag{6.60}$$

Note that the previous integral is performed only up to the Debye frequency, as the phonon density of states is equal to zero beyond that point. Using the change of variable $x = \frac{\hbar\omega}{k_b T}$ (and thus $dx = \frac{\hbar}{k_b T} d\omega$), this equation becomes:

$$E = \frac{3V\hbar}{2\pi^2 v_0^3} \left(\frac{k_b T}{\hbar}\right)^4 \int_0^{\frac{\hbar\omega_D}{k_b T}} \frac{x^3}{e^x - 1} dx \tag{6.61}$$

Let us now make use of the Debye temperature Θ_D defined in Eq. (6.46) and the Debye wavenumber k_D in Eq. (6.47) to express:

$$\frac{1}{(\Theta_D)^3} = \frac{1}{k_D^3} \left(\frac{k_b}{\hbar\nu_0}\right)^3 = \frac{V}{6\pi^2 N} \left(\frac{k_b}{\hbar\nu_0}\right)^3$$

Using Eq. (6.46) for the boundary of the integral, Eq. (6.61) can then be rewritten as:

$$E = 9Nk_b \frac{1}{(\Theta_D)^3} T^4 \int_0^{\frac{\Theta_D}{T}} \frac{x^3}{e^x - 1} dx \tag{6.62}$$

For *high temperatures*, where $k_b T \gg \hbar\omega_D$ or simply $T \gg \Theta_D$, the integral in Eq. (6.62) is evaluated close to zero, i.e., $0 < x < \frac{\Theta_D}{T} << 1$. The function in the integral can thus be approximated as follows:

$$\frac{x^3}{e^x - 1} \approx \frac{x^3}{(1 + x) - 1} = x^2$$

where we have used the approximation $\exp(x) \approx 1 + x$ for $x \to 0$. As a result, Eq. (6.62) becomes successively:

$$\begin{aligned} E &\approx 9Nk_b \frac{1}{(\Theta_D)^3} T^4 \int_0^{\frac{\Theta_D}{T}} x^2 dx \\ &= 9Nk_b \frac{1}{(\Theta_D)^3} T^4 \left[\frac{x^3}{3}\right]_0^{\frac{\Theta_D}{T}} \\ &= 3Nk_b \frac{1}{(\Theta_D)^3} T^4 \left(\frac{\Theta_D}{T}\right)^3 \end{aligned}$$

and finally:

$$E \approx 3Nk_bT \tag{6.63}$$

The heat capacity is thus obtained after differentiating this expression with respect to the temperature as in Eq. (6.63):

$$C_v = \left(\frac{dE}{dT}\right)_v = 3Nk_b \tag{6.64}$$

This relation shows that, for high temperatures, i.e., $T \gg \Theta_D$, the heat capacity is independent of temperature. In fact, this could have been easily calculated using classical theory. Indeed, in classical statistical thermodynamics, each mode of vibration is associated with a thermal energy equal to k_bT. Therefore, for a solid with N atoms, each having three vibrational degrees of freedom, we get *3 N* modes; the total thermal energy is then $3Nk_bT$, as derived in Eq. (6.63); and the heat capacity is found to be equal to Eq. (6.64). This is known as the law of Dulong and Petit, which is based on classical theory. The molar heat capacity, that is, the value of the heat capacity for 1 mole of atoms, is calculated for N equal to the Avogadro number $N_A = 6.02204 \times 10^{23}$ mol^{-1} and is $C_v = 3\mathcal{N}_A k_b = 24.95$ J·mol^{-1}·K^{-1} = 5.96 cal·mol^{-1}·K^{-1}.

This shows that, at high temperatures $T \gg \Theta_D$, the Debye model fits the classical model. For *low temperatures*, however, where $k_bT \ll \hbar\omega_D$ or simply $T \ll \Theta_D$, the heat capacity is not constant with temperature anymore. This is where the quantum theory of phonons is needed and where the accuracy of the Debye model is best appreciated. In this case, the integral in Eq. (6.61) can be extended up to infinity without much error. Moreover, the exponential fraction in the integral can be expressed as:

$$\begin{aligned}\frac{1}{e^x-1} &= \left(\frac{1}{e^x}\right)\left(\frac{1}{1-e^{-x}}\right)\\ &= \frac{1}{e^x}\sum_{n=0}^{\infty}(e^{-x})^n = \sum_{n=1}^{\infty}(e^{-x})^n = \sum_{n=1}^{\infty}e^{-nx}\end{aligned} \tag{6.65}$$

because $x > 0$ and $e^{-x} < 1$. Therefore, the integral in Eq. (6.61) becomes:

$$\begin{aligned}&\int_0^{\frac{\Theta_D}{T}}\frac{x^3}{e^x-1}dx \approx \int_0^{\infty}\frac{x^3}{e^x-1}dx\\ &= \int_0^{\infty}\left(\sum_{n=1}^{\infty}x^3e^{-nx}\right)dx\\ &= \sum_{n=1}^{\infty}\left(\int_0^{\infty}x^3e^{-nx}dx\right)\\ &= \sum_{n=1}^{\infty}I_n\end{aligned} \tag{6.66}$$

where the integral I_n can be simplified after the following successive integration by parts:

$$\begin{aligned} I_n &= \int_0^\infty x^3 e^{-nx} dx \\ &= \left[-x^3 \frac{e^{-nx}}{n}\right]_0^\infty + \int_0^\infty 3x^2 \frac{e^{-nx}}{n} dx && = 0 + \frac{3}{n}\int_0^\infty x^2 e^{-nx} dx \\ &= \frac{3}{n}\left[-x^2 \frac{e^{-nx}}{n}\right]_0^\infty + \frac{3}{n}\int_0^\infty 2x \frac{e^{-nx}}{n} dx && = 0 + \frac{6}{n^2}\int_0^\infty x^2 e^{-nx} dx \\ &= \frac{6}{n^2}\left[-x \frac{e^{-nx}}{n}\right]_0^\infty + \frac{6}{n^2}\int_0^\infty \frac{e^{-nx}}{n} dx && = 0 + \frac{6}{n^3}\int_0^\infty e^{-nx} dx \\ &= \frac{6}{n^3}\left[-\frac{e^{-nx}}{n}\right]_0^\infty \\ &= \frac{6}{n^4} \end{aligned}$$

Thus, Eq. (6.66) can be rewritten as:

$$\int_0^{\frac{\Theta_D}{T}} \frac{x^3}{e^x - 1} dx \approx 6 \sum_{n=1}^{\infty} \frac{1}{n^4} \tag{6.67}$$

The sum in this expression corresponds to $\zeta(4)$, which is called the Riemann zeta function evaluated at 4, and is equal to:

$$\zeta(4) = \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90} \tag{6.68}$$

And Eq. (6.62) becomes:

$$E = 9Nk_b \frac{1}{(\Theta_D)^3} T^4 \frac{6\pi^4}{90}$$

or:

$$E = \frac{3\pi^4}{5} Nk_b \frac{T^4}{(\Theta_D)^3} \tag{6.69}$$

To determine the heat capacity, we must differentiate this expression with respect to the temperature as in Eq. (6.69):

$$C_v = \left(\frac{dE}{dT}\right)_v = \frac{d}{dT}\left(\frac{3\pi^4}{5} N k_b \frac{T^4}{(\Theta_D)^3}\right)$$

or:

$$C_v = \frac{12\pi^4}{5} N k_b \left(\frac{T}{\Theta_D}\right)^3 \tag{6.70}$$

where N is the number of atoms in the crystal. This relation shows that, for low temperatures, i.e., $T << \Theta_D$, the heat capacity is proportional to T^3. The experimentally measured molar heat capacity is shown in Fig. 6.16 for a few solids as a function of temperature.

The figure shows that the Debye model is in good agreement with experimental observations, both in the high-temperature and the low-temperature regions.

Example

Q Calculate the Debye temperature for InP, given that the $v_l = 4.594 \times 10^3$ m·s^{-1}, $v_t = 3.085 \times 10^3$ m·s^{-1}, and the mass density of InP is $d = 4.81 \times 10^3$ kg·m^{-3}.

A We make use of the expression giving the Debye temperature, $\Theta_D = \frac{\hbar \omega D}{k_b}$, where the Debye frequency $\omega_D = v_0 k_D$ is calculated knowing $\frac{1}{v_0^3} = \frac{1}{3}\left(\frac{1}{v_l^3} + \frac{2}{v_t^3}\right)$ and

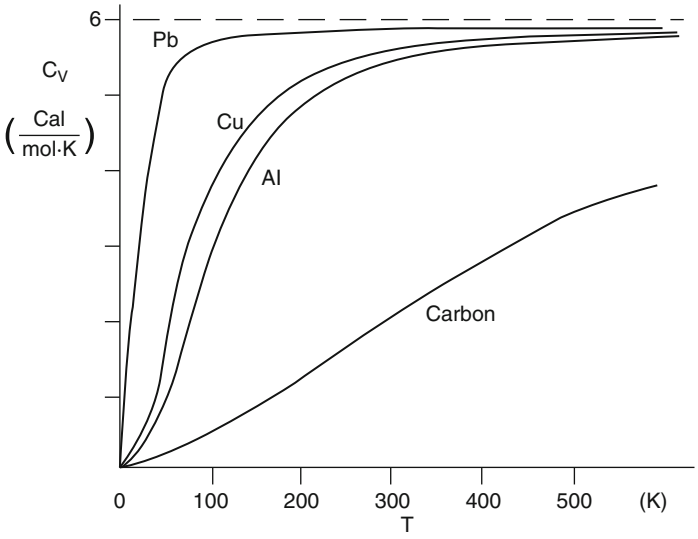


Fig. 6.16 Temperature dependence of the molar heat capacity C_v of some materials. At low temperatures, the heat capacity follows a T^3 relation (Hummel 1993, Fig. 19.1. © 1985, 1993 by Springer-Verlag Berlin Heidelberg. With kind permission of Springer Science and Business Media)

$k_D^3 = \frac{6\pi^2 N}{V} = 6\pi^2\left(\frac{2d}{M_{In}+M_P}\right)$ similarly to the previous example. Numerically, we successively obtain:

$v_0 = \left[\frac{1}{3}\left(\frac{1}{v_l^3}+\frac{2}{v_t^3}\right)\right]^{-1/3} = \left[\frac{1}{3}\left(\frac{1}{\left(4.594\times10^3\right)^3}+\frac{2}{\left(3.085\times10^3\right)^3}\right)\right]^{-1/3}$ or:

$$v_0 = 3.37\times10^3\,\mathrm{m}\cdot\mathrm{s}^{-1}.$$

In addition, we have:

$$k_D = \left(6\pi^2\left(\frac{2\times4.81\times10^3}{(114.8+31)\times1.67264\times10^{-27}}\right)\right)^{1/3}\text{ or}$$

$$k_D = 1.33\times10^{10}\,\mathrm{m}^{-1}$$

which leads to:

$$\omega_D = 4.47\times10^{13}\,\mathrm{Hz}\,\text{and}$$

$$\Theta_D = \frac{\left(1.05458\times10^{-34}\right)\left(4.47\times10^{-13}\right)}{1.38066\times10^{-23}} = 341.5\,K$$

Throughout this discussion, we realized that the Debye temperature Θ_D played a significant role in the heat capacity of a material. It indicates the separation between the high-temperature region where classical theory is valid and the low-temperature region where quantum theory is needed. The Debye temperature can be measured by fitting the experimental data of Fig. 6.16 to Eq. (6.70).

Electronic Contribution to the Heat Capacity

The previous discussion has considered the contribution of lattice vibrations or phonons to the heat capacity. This is valid for dielectric, i.e., insulating, materials. But, unlike dielectric materials, metals have a large number of free electrons, N_f, which can also absorb thermal energy, thus increasing the overall heat capacity of the metal. The contribution of electrons to the total heat capacity, denoted C_v^{el}, can be found as:

$$C_v^{el} = \frac{\pi^2}{2}\frac{N_f k_b^2}{E_F}T \tag{6.71}$$

$$C_v^{el} = \gamma T$$

where N_f is the total number of free electrons in the crystal, E_F is the Fermi energy, k_b the Boltzmann constant, and T the absolute temperature. The mathematical steps involved in the calculation of C_v^{el} are quite challenging and are beyond the scope of

this textbook. Only a few defining equations will be listed here. The heat capacity C_v^{el} is defined by:

$$C_v^{el} = \left(\frac{dE}{dT}\right)_{N_f} \tag{6.72}$$

where E is the energy of all the electrons in the crystal and is given by:

$$E = \int_0^\infty \varepsilon g_{3D}(\varepsilon) f_e(\varepsilon) d\varepsilon \tag{6.73}$$

where $f_e(\varepsilon)$ is the Fermi-Dirac distribution defined in Eq. (5.28) and $g_{3D}(\varepsilon)$ is the three-dimensional electronic density of states of free electrons given by:

$$g_{3D}(\varepsilon) = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{3/2}\sqrt{(\varepsilon)} \tag{6.74}$$

with m^* being the electron effective mass. The temperature dependence of E is included in the Fermi-Dirac distribution function.

We can see from Eq. (6.71) that the electronic contribution C_v^{el} to the heat capacity depends linearly on temperature and thus can be discriminated from the T^3 dependence of the lattice or phonon contribution denoted C_v^{ph} (Eq. 6.70) at low temperatures. It is interesting to consider the ratio of C_v^{el} to C_v^{ph}:

$$\frac{C_v^{el}}{C_v^{ph}} = \frac{\frac{\pi^2}{2}\frac{N_f k_b^2}{E_F}T}{\frac{12\pi^4}{5}Nk_b\left(\frac{T}{\Theta_D}\right)^3} = \frac{5}{24\pi^2}\frac{N_f}{N}\frac{k_b}{E_F}\frac{\Theta_D^3}{T^2} \tag{6.75}$$

where Θ_D is the Debye temperature. By introducing the Fermi temperature T_F such that:

$$E_F = k_b T_F \tag{6.76}$$

And Eq. (6.75) becomes:

$$\frac{C_v^{el}}{C_v^{ph}} = \frac{5}{24\pi^2}\frac{N_f}{N}\frac{\Theta_D^3}{T^2 T_F} \tag{6.77}$$

The ratio $\frac{N_f}{N}$ expresses the average number of free electrons that each atom contributes to the crystal. Equation (6.77) shows that, as the temperature is increased, the contribution of the lattice to the heat capacity exceeds that of electrons. This occurs at a temperature T_0 such that $C_v^{el} = C_v^{ph}$ or:

$$T_0 = \sqrt{\frac{5}{24\pi^2}\frac{N_f}{N}\frac{\Theta_D^3}{T_F}} \tag{6.78}$$

Numerically, one can find that this temperature is only a few percent of the Debye temperature, i.e., a few degrees K (Table 6.1). This means that the contribution of electrons to the heat capacity can only be observed at very low temperatures.

Example

Q Calculate the ratio of ${C_v^{el}}/{C_v^{ph}}$ at 4.2, 30, 77, and 296 K for Cu (assume $\Theta_D = 340\,K$ and $E_F = 7$ eV).

A We start from the expression for the above ratio: $\frac{C_v^{el}}{C_v^{ph}} = \frac{5}{24\pi^2}\frac{N_f}{N}\frac{k_b}{E_F}\frac{\Theta_D^3}{T^2}$. Since Cu has two free electrons per atom, we can write $\frac{N_f}{N} = 2$. This leads to:

$$\frac{C_v^{el}}{C_v^{ph}} = \frac{5}{24\pi^2} \times 2 \times \frac{1.38066 \times 10^{-23}}{7 \times 1.60218 \times 10^{-19}}\frac{340^3}{T^2} = \frac{20.43}{T^2}$$

which gives:

$$\frac{C_v^{el}}{C_v^{ph}} = 1.16\ (4.2\text{K}), 0.023\ (30\text{K}), 0.034\ (77\text{K}), 0.00023\ (296\text{K}).$$

6.2.3 Thermal Expansion

Beside a few notable exceptions, it is commonly known that the volume of a heated solid increases. This phenomenon is called thermal expansion.

If a material of length L is heated through a *small* temperature change ΔT, the change in length ΔL is proportional to the original length and to the change in temperature. The coefficient of linear expansion α_L is called the thermal expansion coefficient and is defined by the following relationship:

$$\frac{\Delta L}{L} = \alpha_L \Delta T \tag{6.79}$$

The linear expansion coefficients of a few solids are shown in Table 6.2.

As Eq. (6.79) describes, an isotropic material exhibits equal thermal expansion in all directions. Some cases in the real world, however, can be more complex than implied by Eq. (6.79). The coefficient α_{L} can vary with temperature, so that the amount of expansion not only depends upon the temperature change but also upon the absolute temperature of the material.

Some materials are not isotropic and have a different value for the coefficient of linear expansion dependent upon the axis along which the expansion is measured.

Table 6.2 Thermal expansion coefficients of a few solids (Chemical Rubber Company 1997; Grigoriev and Meilikhov 1997)

Solid	α_L ($\times 10^{-5}$ K^{-1})
NaCl	3.96
Pb	2.89
Al	2.31
Ag	1.89
Cu	1.65
Au	1.42
Fe	1.18
C (diamond)	1.18
Ordinary glass	0.90
Ge	0.582
GaAs	0.54
InSb	0.47
Si	0.468
AlAs	0.35
Si_3N_4	0.27
Pyrex glass	0.32
Invar	0.07
Quartz glass	0.05

For instance, with increasing temperature, calcite ($CaCO_3$) crystals expand along one crystal axis and contract ($\alpha_L < 0$) along another axis.

Engineers in the semiconductor field are often extremely concerned about the thermal expansion rate of a material when designing a device or system that must operate over a range of temperatures. Improperly packaging a semiconductor device without giving careful consideration to the thermal expansion properties of the materials can result in reliability problems and reduced lifetime of the device. As a result, most companies perform thermal cycling tests of their devices to determine whether or not thermal expansion is a possible failure mechanism.

The problems associated with thermal expansion are most severe when two materials of different thermal expansion coefficients are permanently bonded together, such as in integrated circuits. For example, if the thermal expansion properties of a metal heat sink are not properly matched to the thermal expansion properties of the semiconductor material, the brittle semiconductor can crack as the device is heated and cooled. In fact, copper and other metals exhibit thermal expansion properties that are an order of magnitude greater than that of semiconductors such as Si and GaAs, making it very problematic to attach these materials directly. In order to address this issue, many semiconductor devices are packaged using intermediate die attachment materials as well as advanced solder alloys and optimized package materials as illustrated in Fig. 6.17. Some examples of advanced packaging processes that rely on optimizing the coefficient of thermal expansion are high-power RF-electronics and lasers (Fig. 6.17).

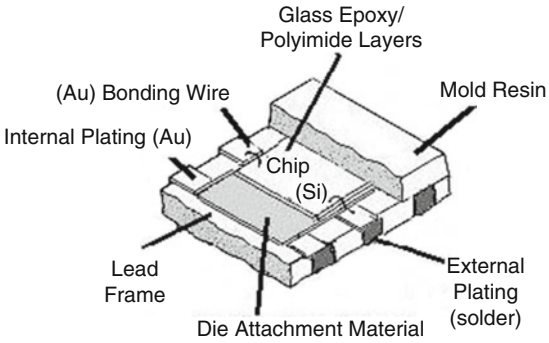
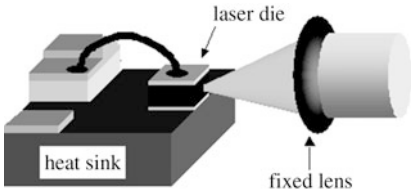


Fig. 6.17 Cutaway illustration of an advanced semiconductor device package. To avoid cracking and stresses and for devices where alignment is critical, packaging materials must be chosen with compatible thermal expansion coefficients

Example

Q A semiconductor laser is affixed to a copper heat sink and sealed into a package inside of a factory clean room environment where the ambient temperature is 20 °C. The lasers are then installed in scientific equipment monitoring gas emissions from a volcano in Hawaii.



The package also contains a collimating lens that is fixed in place and aligned with the central axis of the laser beam. If the ambient temperature in Hawaii is48 °C, how far off axis will the laser be when the device is in operation? (Assume that thermal expansion has a negligible effect on the in-plane expansion of the heat sink. Also assume that the heat sink is 3 mm long on each side and 1 mm tall).

A Equation (6.79) describes the linear expansion of a material: $\frac{\Delta L}{L} = \alpha_L \Delta T$. Cu has a coefficient of linear expansion, α_L, equal to $1.67 \times 10^{-5}\ \mathrm{K}^{-1}$. The heat sink is originally 1mm tall (L), and the temperature difference, ΔT, is equal to 48 °C−20 °C = 28 °C = 28 K.

Thus, the change in length of the heat sink is equal to:$\Delta L = (1.67 \times 10^{-5} K^{-1})(28K)(1\ mm) = 4.68 \times 10^{-4} mm$
or 0.468 μm.

Thermal expansion means that the average distance between atoms increases when the temperature goes up and is therefore related to atomic vibrations or

phonons in a solid. It can be easily understood that at a higher temperature, the atomic vibrations will be more intense, the distances between atoms will be higher, and therefore the overall solid volume will be larger. The mathematical treatment of this relationship is beyond the scope of the discussion. We will merely give a brief and simple description of the phenomenon.

We saw in Sect. 6.1.2 that the *equilibrium* interatomic distance $r = R_0$ is determined by the minimum of the atomic interaction potential energy $U(r)$. In thermodynamics, for such a system at thermal equilibrium at a temperature T, the *average* interatomic distance is denoted <R> and is given by the Maxwell-Boltzmann distribution:

$$\langle R\rangle = \frac{\int_{-\infty}^{\infty} Re^{-\frac{U(R)}{k_bT}}dR}{\int_{-\infty}^{\infty} e^{-\frac{U(R)}{k_bT}}dR} \tag{6.80a}$$

By introducing the displacement $x = R - R_0$ and expressing $U(R)$ as a function of x as was done in Sect. 6.1.2 (Eq. 6.2), we can rewrite this equation as:

$$\langle R\rangle = \frac{\int_{-\infty}^{\infty} (R_0 + (R - R_0))e^{-\frac{U(R)}{k_bT}}dR}{\int_{-\infty}^{\infty} e^{-\frac{U(R)}{k_bT}}dR} = R_0\frac{\int_{-\infty}^{\infty} e^{-\frac{U(R)}{k_bT}}dR}{\int_{-\infty}^{\infty} e^{-\frac{U(R)}{k_bT}}dR} + \frac{\int_{-\infty}^{\infty} xe^{-\frac{U(x)}{k_bT}}dx}{\int_{-\infty}^{\infty} e^{-\frac{U(x)}{k_bT}}dx}$$

or:

$$\langle R\rangle = R_0 + \frac{\int_{-\infty}^{\infty} xe^{-\frac{U(x)}{k_bT}}dx}{\int_{-\infty}^{\infty} e^{-\frac{U(x)}{k_bT}}dx} \tag{6.80b}$$

For low temperatures and thus small vibrational amplitudes ($x << R_0$), one can approximate the potential energy $U(x)$ with terms up to the second order in x (i.e., x^2) as was done in Eq. (6.3). This is the harmonic approximation. In this case, the exponential $e^{-\frac{U(x)}{k_bT}}$ is an even function of x, $xe^{-\frac{U(x)}{k_bT}}$ is an odd function of x, and therefore $\int_{-\infty}^{\infty} xe^{-\frac{U(x)}{k_bT}}dx = 0$ and $\langle R\rangle = R_0$. This means that, in the harmonic case, the average interatomic distance <R> is exactly R_0, the distance corresponding to the potential energy minimum.

At higher temperatures, the atomic displacement x is large enough so that higher order terms in Eq. (6.2) need to be included (e.g., x^3), causing anharmonic effects. In this case, the exponential $e^{-\frac{U(x)}{k_bT}}$ is not an even or odd function of x anymore, and the integral fraction in Eq. (6.38) is strictly positive. As a result, $\langle R\rangle > R_0$ which means

that the average interatomic distance becomes larger than R_0, i.e., there is thermal expansion. We see that thermal expansion is a direct result of anharmonic effects in the atomic interaction potential.

6.2.4 Thermal Conductivity

In the previous few sections, we saw that a lattice could receive and store thermal energy, heat through lattice vibrations, i.e., by creating more phonons, or through free electrons in a metal by gaining more kinetic energy. The lattice vibrations generate waves that can propagate, while free electrons can move in a metal. The thermal energy can thus be transported from one end of the solid to another. This characteristic is called thermal conductivity and is also an important parameter when designing a device or system.

Depending on the thermal conductivity of the materials used, heat may build up from the operation of the device and lead to failure of the device or system. Removal of excess heat has become a very critical issue in semiconductor design in recent years, especially in the design of modern high-density computer chips and high-power optoelectronic semiconductors. In the semiconductor industry, Moore's law has predicted that the number of transistors on a chip doubles every 18 months. This has led to both a reduction of the size of transistors and an increase in the packing density. The increase in transistor density has also lead to a significant increase in the power density (heat) in the same area that needs to be removed from the chip.

The thermal conductivity of a solid is quantified through a positive parameter called the thermal conductivity coefficient K (read "kappa") which is defined as:

$$J_T = -\kappa \frac{dT}{dx} \tag{6.81}$$

where J_T is the thermal current density, i.e., the thermal energy transported across a unit area per unit time. This is expressed in units of $\mathrm{J{\cdot}cm^{-2}{\cdot}s^{-1}}$ or $\mathrm{W{\cdot}cm^{-2}}$. $\frac{dT}{dx}$ is the temperature gradient, which is the rate at which the temperature changes from one region of the solid to another. The thermal conductivity coefficient thus has the units of $\mathrm{W{\cdot}cm^{-1}{\cdot}K^{-1}}$ (or $\mathrm{W{\cdot}m^{-1}{\cdot}K^{-1}}$). Values of the thermal conductivity of a few materials are given below in Table 6.3 and Fig. A. in Appendix A.3.

Equation (6.81) expresses that there is a flux of thermal energy within the solid as a result of a difference of temperature between two regions. The minus sign means that the thermal energy flows from the higher-temperature region to the lower-temperature region. This relation is analogous to the electrical current which originates from a difference in electrical potential. In Eq. (6.39), we assumed that the thermal current and the temperature gradient occurred along one direction. In a three-dimensional case, the current and the gradient would be simply replaced by vectors. The simplification here does not reduce the generality of the physical concepts which will be derived. Moreover, in this section, we will only be interested

Table 6.3 Thermal conductivities of a few solids (Chemical Rubber Company 1997; Adachi 2004)

Solid	κ (W·m^{-1}·K^{-1})
Pyrex glass	1.1
NaCl	6.4
Pb	35
GaAs	56
Ge	64
GaP	77
Fe	80
AlN	82
InP	68
Si	124
BeO	210
Al	237
Au	317
Cu	401
Ag	429
C (diamond)	1000

in the qualitative properties of the thermal conductivity. An exhaustive mathematical treatment can therefore be avoided.

Copper has become the material of choice for most heat-spreading applications in microelectronics because it is a material with one of the highest thermal conductivities and affordable costs. In some cutting edge devices, however, even copper is falling short of adequately removing heat from semiconductor devices, and the engineers and materials scientists have had to think of alternative approaches. One such approach has been to use diamond because it has a thermal conductivity several times larger than that of copper. Commercial manufacturing of diamond heat-spreading materials through the use of chemical vapor deposition (CVD) has reduced the material's cost and improved availability and made diamond heat spreaders a viable solution for high-heat load applications, such as power laser diodes.

Thermal conductivity can be viewed as the result of phonons (quasiparticle) moving from a hotter to a colder region and undergoing collisions with one another or against material imperfections (defects, boundaries) so that their energy can be transferred in space. These collisions are also often referred to by using the more general term scattering. The mathematical model commonly followed makes use of the kinetic theory of gases, in which (i) each quasiparticle is modeled as a free moving particle in space with a momentum and an energy, (ii) each quasiparticle is subject to instantaneous collision events with other particles, (iii) the probability for a collision to occur during an interval of time dt is proportional to dt, and (iv) the particles reach thermal equilibrium only through these collisions.

Similar to the heat capacity, there are two contributions to the thermal conductivity: a lattice contribution (phonons) denoted κ_{ph} and an electronic contribution (electrons) denoted κ_e.

The lattice contribution κ_{ph} can be regarded as the thermal conductivity of a phonon gas. Using the kinetic theory of gases, the following expression can be derived for the lattice contribution:

$$\kappa_{ph} = \frac{1}{3}\left(\frac{C_v^{ph}}{V}\right) v_0 \Lambda \tag{6.82}$$

where $\left(\frac{C_v^{ph}}{V}\right)$ is the heat capacity per unit volume of the solid considered and v_0 is the average phonon velocity. The parameter Λ is the mean free path of a phonon between two consecutive collisions and is central to the thermal conductivity process.

There are two types of phonon-phonon interactions in crystals. The first one involves what is called normal processes which conserve the overall phonon momentum, $\vec{k}_1 + \vec{k}_2 + \vec{k}_3 = 0$, but not phonon number (phonons are bosons and are not subject to particle number conservation) where $\vec{k}_1$, $\vec{k}_2$, and $\vec{k}_3$ are the momenta of three interacting phonons. The second type is called umklapp processes and is such that $\vec{k}_1 + \vec{k}_2 + \vec{k}_3 = n\vec{K}$, where $n = 1, 2, 3\ldots$ is an integer and $\vec{K}$ is a reciprocal lattice vector. We recall from Chaps. 4 and 5 that electron and lattice momentum in a crystal is only conserved give or take a reciprocal lattice vector. Equation (6.40) was first applied by Debye to describe thermal conductivity in dielectric (insulating) solids.

At very low temperatures, i.e., $T \ll \Theta_D$, the average number of phonons given in Eq. (6.33) tends toward zero. The phonon-phonon scattering becomes negligible, and the mean free path Λ is determined by the scattering of phonons against the solid imperfections or even the solid boundaries. Λ thus increases until it is equal to the geometrical size of the sample. Then, the thermal conductivity behaves as the heat capacity C_v^{ph} and has a T^3 dependence (Eq. (6.80)). In particular, $\kappa_{ph} \to 0$ when $T \to 0$. These are shown in Fig. 6.18a for Λ and Fig. 6.18b for κ_{ph}.

For higher temperatures, i.e., $T \gg \Theta_D$, we saw in Sect. 6.1.6 that the average number of phonons is proportional to T. Thus, phonon-phonon interactions become increasingly dominant as the temperature increases. Since the collision frequency should be proportional to the number of phonons with which a phonon can collide, Λ ends up being proportional to $1/T$ at higher temperatures, as shown in Fig. 6.18a. At the same time, we saw that in the heat capacity C_v^{ph} saturates at high temperatures (Eq. 6.71). The thermal conductivity κ_{ph} therefore has a $1/T$ dependence in this regime, as shown in Fig. 6.18b.

Another contribution to the thermal conductivity arises from electrons and mainly concerns metals which have a large concentration of free electrons. Here, again, the kinetic theory of gases leads to an expression of the electronic contribution κ_{el} similar to Eq. (6.82):

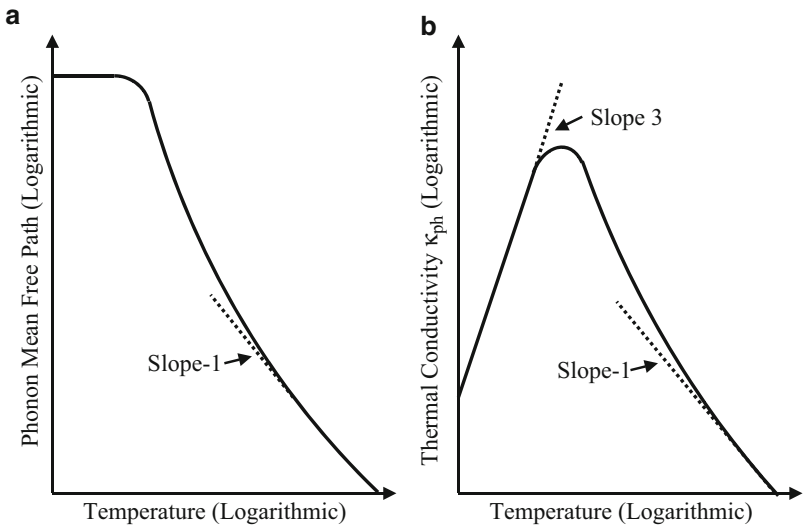


Fig. 6.18 Variation of (**a**) phonon mean free path and (**b**) lattice thermal conductivity as a function of temperature. At low temperatures, as the phonon-phonon interaction and scattering decrease, the phonon mean free path is determined by crystal imperfections which are independent of temperature, and the thermal conductivity follows a T^3 dependence. At high temperatures, phonon-phonon scattering increases, and both the phonon mean free path and the thermal conductivity decrease as T^{-1}

$$\kappa_{el} = \frac{1}{3}\left(\frac{C_v^{el}}{V}\right) v_e \Lambda_e \tag{6.83a}$$

where $\left(\frac{C_v^{el}}{V}\right)$ is the electronic contribution to the heat capacity per unit volume of the solid considered and v_e is the average electron velocity. The parameter Λ_e is the mean free path of an electron and describes how far an electron can travel on average between two consecutive collisions. We will not in this chapter discuss the various scattering mechanisms for an electron because of their large number and complexity. Electronic transport and relaxation times will be discussed in more details in Chap. 8. An interesting relationship can be derived linking the thermal conductivity and electrical conductivity (σ_{el}) of the free electron gas using Eqs. 6.71 and 6.83. This is known as the Wiedemann-Franz law and can be written as:

$$\kappa_{el} = \frac{\pi^2 k_b^2}{3q^2} T \sigma_{el} \tag{6.83b}$$

The electrical conductivity σ_{el} has not yet been discussed and is treated in detail in Chap. 8, Sect. 8.2. It is measured in units of siemens/m or S/m.

We will conclude by providing a numerical estimate of this contribution and compare it to the lattice contribution. At room temperature, on the one hand, a

typical phonon has a mean free path of 3×10^{-6} cm, a velocity of 10^5 cm·s^{-1}, and a heat capacity of 25 J·K^{-1}·mol^{-1}, yielding a thermal conductivity of $\kappa_{ph} \approx 2.5$ W·cm^{-1}·K^{-1}. On the other hand, for a pure (perfect) metal, an electron has a mean free path of 10^{-5} cm, a velocity of 10^8 cm·s^{-1}, and a heat capacity of 0.5 J·K^{-1}·mol^{-1}, yielding a thermal conductivity of $\kappa_{ph} \approx 250$ W·cm^{-1}·K^{-1}. This clearly shows that the electrons in a pure metal are responsible for almost all the heat transfer. However, if the metal has many defects, the phonon contribution may be comparable with the electron contribution.

6.2.5 Summary

In this chapter, we have shown that phonons in solids are responsible for important contributions to the thermal properties of crystals. This includes heat capacity, thermal expansion, and thermal conductivity. The Debye model of phonons was presented, and it was shown that, despite the considerable simplifications made to the spectrum, the model still accurately describes the temperature dependence of the heat capacity and the thermal conductivity coefficients as measured experimentally in crystals. The subject of thermal conductivity has acquired more importance recently in view of the work on thermoelectricity and heat energy harvesting. Thermal transport and how to control it are treated in detail in Chap. 12 of this book.

Problems for Phonons and Thermal Properties

1. Explain why there is no optical phonon in the dispersion curve for the one-dimensional monatomic chain of atoms.
2. Explain why there is a forbidden range of vibration energies between the optical and acoustic phonon branches. Solve Eq. (6.22) for the case when $k = \pi/a$.
3. The one-dimensional monatomic harmonic crystal (Sect. 6.1.3) is in fact a particular case of the diatomic model described in Sect. 6.1.4, for which the two atoms are identical. To prove this, show that the expression for the diatomic harmonic crystal can be transformed into an expression similar to the monatomic crystal. Solve Eq. (6.22) in the limit $M_1 = M_2 = M$. What considerations do you have to take into account to do this?
4. In the chapter, the phonon frequencies at the center of the zone $k = 0$ were determined for the diatomic molecule. Calculate the phonon frequencies at the zone boundary $k = \pi/a$.
5. Plot the shapes of the optical and acoustic branches in the dispersion relation for four different ratios of masses: $\frac{M_1}{M_2} = 10$, 5, 2, and 1. Show that, in the case of two identical atoms, there is actually only one acoustic branch and no optical branch for the dispersion relation.
6. In Sect. 6.1.4, we calculated the ratio of the displacement amplitudes A and B for the long wave limit ($k \rightarrow 0$) for both the optical and acoustic phonon branches

and then determined the displacement of the atoms with respect to each other. Calculate Eq. (6.26), the ratio of the displacement amplitudes, in the short wave limit ($k \rightarrow \pi/a$), and draw the displacement of the atoms with respect to each other.

7. Suppose that a light wave of wavelength 3 μm is absorbed by a one-dimensional diatomic harmonic chain with atoms of mass 4×10^{-26} kg and 5×10^{-26} kg and atomic spacing of 4.5 Å. What is the force constant in MKS units?
8. From the figures for the phonon dispersion curves for Si and GaAs plus the equations for optical and acoustic phonons, explain why the energy for the Si curves is higher in energy than the curves for GaAs? Assume that the elastic constant is about the same for both materials. Also, why do the optical and acoustic phonon branches cross at the zone boundary for Si but not for GaAs?
9. Plot the average number of phonons $N(\omega) = \frac{1}{\exp\left(\frac{\hbar\omega}{k_b T}\right) - 1}$ for at least five values of T to show its evolution with increasing temperatures. For each one, plot the function $F(\omega) = \frac{k_b T}{\hbar\omega}$, and show that it is a good approximation for $N(\omega)$ for high temperatures, i.e., $k_b T >> \hbar\omega$.
10. Let us model a rigid bar as a linear monatomic chain of atoms, as in Sect. 6.1.3 with the same notations. We further assume that the equilibrium interatomic separation is a and that its cross section is a^2. Its Young's modulus E_Y is defined as the ratio of the stress applied in one direction divided by the relative elongation in this same direction. The stress is the ratio of the interatomic force ($F_{n,n-1}$) divided by the cross-sectional area (a^2) on which this force is applied. The relative elongation is the interatomic displacement divided by the equilibrium separation. The Young's modulus has the dimension of a pressure and is expressed in Pa (Pascal). The solid density M_V is the ratio of the mass of the solid to its volume. Here, we assume that the mass of an atom is M and that there is only one atom in a volume of a^3.

    Show that the sound velocity, defined in Sect. 6.1.7, is equal to the ratio: $\sqrt{\frac{E_Y}{M_V}}$.
11. From the speed of sound equation, $\nu = (B/\rho)^{1/2}$, calculate the speed of sound in silicon and compare with the speed of sound in gallium arsenide. Assuming that the largest effect on the velocity comes from the density, why is this result expected?

Problems for Thermal Properties of Crystals

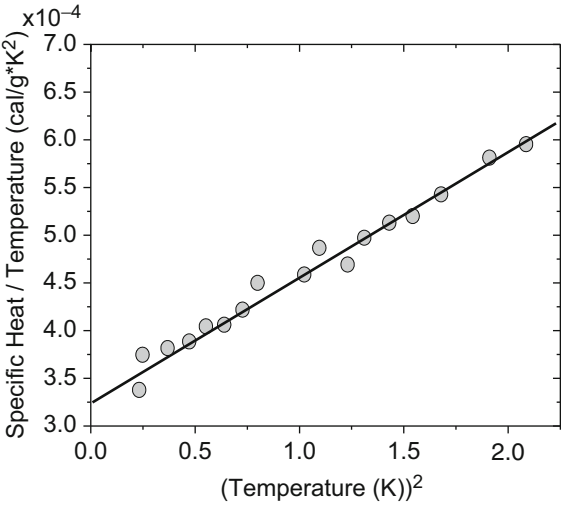
1. In your own words, describe the meaning of the phonon density of states.
2. In your own words, describe the meaning of the Debye frequency and the Debye temperature. Develop a simple equation relating the Debye frequency, Debye temperature, and Debye wavelength.
3. Determine the Debye temperature Θ_D, Debye wavelength, and the Debye frequency ω_D for diamond given that the lattice constant for this material is

3.56 Å, the density of diamond is 3.52×10^3 kg·m^{-3}, and the speed of sound in diamond is 12,000 m·s^{-1}.

4. In your own words, describe the meaning of heat capacity. How is heat capacity related to specific heat?
5. Starting from the expression of the total energy carried by the lattice vibrations in Eq. (6.60), show that the heat capacity $C_v = \left(\frac{dE}{dT}\right)_v$ can be written as:

$$C_v = 9Nk_b\left(\frac{T}{\Theta_D}\right)^3 \int_0^{\frac{\Theta_D}{T}} \frac{x^4 e^x}{(e^x - 1)^2} dx$$

6. It takes 450 cal to raise the temperature of a metallic sample from 20 to 35 °C. What is the heat capacity of the metal sample? If the sample has a mass of 78 g, what is the specific heat of the sample?
7. The specific heat of metals is dominated by the electronic contribution at low temperatures and by phonons at high temperatures. At what temperature are the two contributions equal in rubidium? Note that $\gamma = 2.41$ mJ/(mole K^2) for rubidium. Briefly describe your thinking.
8. The figure below illustrates measurements of the specific heat (plotted as C/T versus T^2) for a crystalline element. Use what you know about the origins and temperature dependence of the specific heat capacity to determine whether the element is Na or Si. Discuss both possibilities.



Experimental data of the specific heat of an unknown element.

9. In your own words, describe the meaning of thermal expansion in solid-state engineering.
10. Look up in tables or reference books the room temperature lattice constants for the following crystals: aluminum, copper, iron, silicon, germanium, and diamond. Using the coefficients of linear expansion, plot the values of the lattice constants up to a temperature of 1000 °C.
11. In your own words, briefly describe the meaning of thermal conductivity and the physical processes that influence the thermal conductivity.
12. Diamond is an electrical nonconductor; however, the thermal conductivity of diamond is greater than the thermal conductivity of copper for $T > 40$ K. How can this be explained?

References

Adachi S (2004) Handbook on physical properties of semiconductors Volume 2-III-V compound semiconductors. Kluwer Academic, Boston
Chemical Rubber Company (1997) CRC handbook of chemistry and physics. CRC Press, Cleveland
Grigoriev IS, Meilikhov EZ (1997) CRC handbook of physical quantities. CRC Press, Boca Raton
Hummel RE (1993) Electronic properties of materials. Springer, New York, p 335
Dolling G (1963) Lattice vibrations in crystals with the diamond structure. In: Inelastic scattering of neutrons in liquids and solids, vol 2. International Atomic Energy Agency, Vienna, p 41
Waugh JLT, Dolling G (1963) Crystal dynamics of gallium arsenide. Phys Rev 132:2411

Further Reading

Ashcroft NW, Mermin ND (1976) Solid state physics. Holt, Rinehart and Winston, New York
Cochran W (1973) The dynamics of atoms in crystals. Edward Arnold Limited, London
Cohen MM (1972) Introduction to the quantum theory of semiconductors. Gordon and Breach, New York
Ferry DK (1991) Semiconductors. Macmillan, New York
Ibach H, Lüth H (1990) Solid-state physics: an introduction to theory and experiment. Springer, New York
Kasap SO (1997) Principles of engineering materials and devices. McGraw-Hill, New York
Kittel C (1976) Introduction to solid state physics. Wiley, New York
Maxwell JC (1952) Matter and motion. Dover, New York
Peyghambarian N, Koch SW, Mysyrowicz A (1993) Introduction to semiconductor optics. Prentice-Hall, Englewood Cliffs
Reissland JA (1973) Physics of phonons. Wiley, London
Sapoval B, Hermann C (1995) Physics of semiconductors. Springer, New York
Ashcroft NW, Mermin ND (1976) Solid state physics. Holt, Rinehart and Winston, New York
Born M, Huang K (1954) Dynamical theory of crystal lattices. Clarendon Press, Oxford
Cochran W (1973) The dynamics of atoms in crystals. Edward Arnold Limited, London
Cohen MM (1972) Introduction to the quantum theory of semiconductors. Gordon and Breach, New York
Ibach H, Lüth H (1990) Solid-state physics: an introduction to theory and experiment. Springer, New York
Kasap SO (1997) Principles of engineering materials and devices. McGraw-Hill, New York

Kittel C (1976) Introduction to solid state physics. Wiley, New York
Ferry DK (1991) Semiconductors. Macmillan, New York
Maxwell JC (1952) Matter and motion. Dover, New York
Peyghambarian N, Koch SW, Mysyrowicz A (1993) Introduction to semiconductor optics. Prentice-Hall, Englewood Cliffs
Sapoval B, Hermann C (1995) Physics of semiconductors. Springer, New York

Equilibrium Charge Carrier Statistics in Semiconductors

7

7.1 Introduction

In Chap. 4, we discussed the quantum mechanical states of electrons in a periodic crystal potential and the resulting formation of energy bands. We also introduced the concept of effective mass, that of holes, and the Fermi energy which provides an easy way to differentiate a semiconductor from a metal.

In semiconductor devices, most of the properties of interest have their origins in the electrons in the conduction band and the holes in the valence band. Two major functions are important in understanding the behavior of these electrons and holes: the density of states and the Fermi-Dirac distribution function, both of which have been discussed in Chaps. 4 and 5. In this chapter, we will establish the basic relations and formalism for the distribution of electrons in the conduction band and holes in the valence band at thermal equilibrium. We will also introduce the notion of doping and extrinsic semiconductors, in contrast to pure or intrinsic semiconductors.

7.2 Density of States

In Chap. 5, we calculated the density of states of electrons of the conduction band in a three-dimensional semiconductor to be:

$$g_c(E) = \frac{V}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(E - E_C)^{1/2} \tag{7.1}$$

where m_e is the electron effective mass in the conduction band, E_C is the bottom of the conduction band, and V is the volume of the crystal considered. The subscript "c" in g_c indicates that we are considering the conduction band. This expression was calculated for a single band minimum and is valid for direct-gap semiconductors, such as GaAs, where the conduction band minimum occurs at the zone center.

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_7

However, in the case of many other semiconductors, one has to take into account the degeneracy or number g_d of equivalent conduction band minima in the first Brillouin zone.

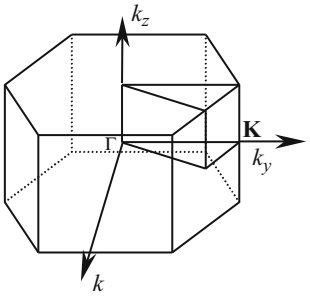
For example, we saw in Fig. 5.17a that the conduction band minimum in Ge occurred along the <111> direction. As there are eight equivalent <111> directions, there are eight equivalent conduction band minima in Ge. However, because the minima occur exactly at the boundary of the first Brillouin zone, each minimum is shared with two neighboring zones and therefore only contributes one half to the density of states. Thus $g_{deg} = 4$, i.e., the expression in Eq. (5.52) needs to be multiplied by a factor 4. In addition, we also saw in Fig. 5.17b that the conduction band minimum in Si occurs at $k \approx 0.8(2\pi/a)$ in the first Brillouin zone along the <100> direction. Since the <100> direction has a sixfold symmetry, this gives rise to six equivalent conduction band minima within the first Brillouin zone, and $g_d = 6$ because the minimum is strictly inside the first Brillouin zone. The expression in Eq. (7.1) then needs to be multiplied by 6. Finally, for GaAs, as shown in Fig. 5.17c, the conduction band minimum occurs at the zone center, and the expression in Eq. (7.1) remains unchanged, i.e., $g_d = 1$.

In other words, the full density of states of electrons in the conduction band is ($E > E_C$):

$$g_c(E) = \frac{V}{2\pi^2} g_d \left(\frac{2m_e}{\hbar^2}\right)^{3/2} (E - E_C)^{1/2} \tag{7.2}$$

Example

Q GaN has the wurtzite crystal structure. The first Brillouin zone is shown in the figure below. From the calculation of the band structure of GaN, it can be seen that there is a shallow conduction band minimum at the symmetry point *K* in the reciprocal lattice. To calculate the density of states given by the expression $g_c(E) = \frac{V}{2\pi^2} g_d \left(\frac{2m_e}{\hbar^2}\right)^{3/2} (E - E_C)^{1/2}$, what is the degeneracy factor g_d which should be used?



A The point K is equally shared by three adjacent Brillouin zones. Because the first Brillouin zone has sixfold symmetry, there are six equivalent points K in the zone. This leads to a total degeneracy of : $6 \times \frac{1}{3} = 2$.

The value of the electron effective mass m_e was determined in Eq. (5.27), in the simple case of a one-dimensional crystal, as the curvature of the conduction band or, in other words, the second derivative of the energy spectrum $E(k)$ such that $E(k)$ can be approximated as:

$$E(k) \approx \frac{\hbar^2}{2m_e} k^2 \tag{7.3}$$

In the more general case of a three-dimensional crystal, the effective mass is a 3×3 matrix, and each element is a function of the direction in which the two derivatives of the energy spectrum $E(\vec{k})$ are performed, k_x, k_y, or k_z.

If the energy spectrum can be approximated as:

$$E(\vec{k}) \approx \frac{\hbar^2}{2} \left(\frac{k_x^2}{m_{xx}} + \frac{k_y^2}{m_{yy}} + \frac{k_z^2}{m_{zz}} \right) \tag{7.4}$$

where m_{xx}, m_{yy}, and m_{zz} correspond to the values of the second partial derivatives in the k_x, k_y, and k_z directions, respectively; then the electron effective mass m_e that is considered in Eq. (7.3) is the average of these three masses and is given by:

$$m_e = \left(m_{xx} m_{yy} m_{zz} \right)^{1/3} \tag{7.5}$$

In the particular case when the energy spectrum can be approximated as:

$$E(\vec{k}) \approx \frac{\hbar^2}{2} \left(\frac{\left(k_x^2 + k_y^2 \right)}{m_t} + \frac{k_z^2}{m_l} \right) \tag{7.6}$$

where m_t and m_l are customarily called the transverse electron effective mass and the longitudinal electron effective mass, respectively; then the electron effective mass m_e that is considered in Eq. (7.5) is the average of these three masses and is given by:

$$m_e = \left(m_t^2 m_l \right)^{1/3} \tag{7.7}$$

A similar relation can be obtained for the electronic density of states in the valence band ($E_V < E$):

$$g_v(E) = \frac{V}{2\pi^2} \left(\frac{2m_h}{\hbar^2} \right)^{3/2} (E_V - E)^{1/2} \tag{7.8}$$

where m_h is the hole effective mass which accounts for the curvature of the valence band and E_V is the top of the valence band. In this expression, there is no degeneracy factor from crystal symmetry because the top of the valence band is unique and always occurs at the center of the first Brillouin zone.

We saw in Sect. 5.4 that the valence band of a semiconductor is composed of two main subbands, the heavy-hole and light-hole bands, each with a different curvature and thus with their own hole effective masses: m_{hh} and m_{lh}, for the heavy-hole effective mass and light-hole effective mass, respectively. As a result, the hole effective mass m_h that is considered in Eq. (7.7) is the following average of these two masses:

$$m_h = \left(m_{hh}^{3/2} + m_{lh}^{3/2}\right)^{2/3} \tag{7.9}$$

7.3 Effective Density of States (Conduction Band)

As discussed in Sub-sect. 5.2.8, the density of states merely provides information about the allowed energy states. To obtain the concentration of electrons in the conduction band, we must multiply this density of states with the Fermi-Dirac distribution (Eq. (5.28)) which gives the probability of occupation of an energy state:

$$n = \frac{1}{V}\int_{E_C}^{\infty} g_c(E) f_e(E) dE \tag{7.10}$$

Expanding this expression using Eq. (5.52) and Eq. (5.28), we get:

$$n = \frac{g_d}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\int_{E_c}^{\infty}\frac{(E-E_C)^{1/2}}{\exp\left(\frac{E-E_F}{k_bT}\right)+1}dE \tag{7.11}$$

Making the change of variable $y = \frac{E-E_C}{k_bT}$, and thus $dy = \frac{1}{k_bT}dE$, the previous integral becomes:

$$\int_{E_C}^{\infty}\frac{(E-E_C)^{1/2}}{\exp\left(\frac{E-E_F}{k_bT}\right)+1}dE = (k_bT)^{3/2}\int_0^{\infty}\frac{y^{1/2}}{\exp\left(y-\frac{E_F-E_C}{k_bT}\right)+1}dy \tag{7.12}$$

We can define the Fermi-Dirac integral as in Eq. (5.56):

$$F_{\frac{1}{2}}(x) = \frac{2}{\sqrt{\pi}}\int_0^{\infty}\frac{y^{1/2}}{1+\exp(y-x)}dy \tag{7.13}$$

using:

$$x = \frac{E_F - E_C}{k_b T} \tag{7.14}$$

Equation (7.12) can be rewritten as:

$$\int_{E_C}^{\infty} \frac{(E - E_C)^{1/2}}{\exp\left(\frac{E-E_F}{k_b T}\right) + 1} dE = (k_b T)^{3/2} \frac{\sqrt{\pi}}{2} F_{\frac{1}{2}}\left(\frac{E_F - E_C}{k_b T}\right) \tag{7.15}$$

and therefore Eq. (7.11) becomes:

$$n = \frac{g_d}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2} (k_b T)^{3/2} \frac{\sqrt{\pi}}{2} F_{\frac{1}{2}}\left(\frac{E_F - E_C}{k_b T}\right) \tag{7.16}$$

Remembering that $\hbar = \frac{h}{2\pi}$, this can be simplified as:

$$n = 2g_d\left(\frac{2\pi k_b T m_e}{h^2}\right)^{3/2} F_{\frac{1}{2}}\left(\frac{E_F - E_C}{k_b T}\right) \tag{7.17}$$

or:

$$n = N_c F_{\frac{1}{2}}\left(\frac{E_F - E_C}{k_b T}\right) \tag{7.18}$$

with:

$$N_c = 2g_d\left(\frac{2\pi k_b T m_e}{h^2}\right)^{3/2} \tag{7.19}$$

N_c is called the effective conduction band density of states. The Fermi-Dirac integral defined in Eq. (7.13) is often approximated with simpler expressions. One commonly encountered situation is when $E_C - E_F >> k_b T$. A semiconductor in this situation is called a non-degenerate semiconductor. Let us give a numerical example. At room temperature ($T = 300$ K), we have $k_b T = 25.9$ meV. Therefore, we can consider that we are in the presence of a non-degenerate semiconductor when the Fermi energy E_F is away from the bottom of the conduction band E_C by a few times 25.9 meV. This is illustrated in Fig. 7.1a. For most of the practical calculations, a distance of $3k_b T$ or more, i.e., $E_C - E_F \geq 3k_b T$, is sufficient.

This approximation means that the Fermi energy is rather far from the bottom of the conduction band and inside the bandgap and that $x << -1$ in Eq. (7.13). Therefore, the exponential function dominates in the denominator for all positive values of $y > 0$, i.e., $1 + \exp(y - x) \approx \exp(y - x)$. Thus:

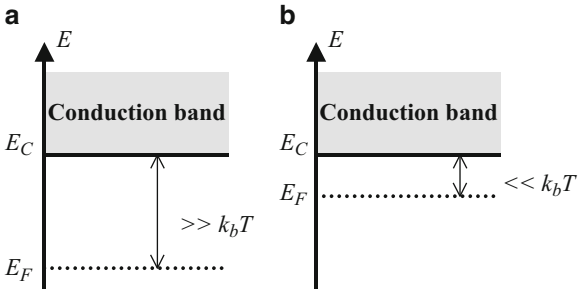


Fig. 7.1 Illustration of the position of the Fermi level with respect to the conduction band (**a**) in a non-degenerate n-type semiconductor: The Fermi energy is far from the edge of the conduction band. (**b**) In a degenerate semiconductor *n*-type semiconductor, the Fermi energy is close to the edge of the conduction band

$$F_{\frac{1}{2}}(x) \approx \frac{2}{\sqrt{\pi}} \int_0^\infty \frac{y^{1/2} dy}{\exp(y-x)} = \frac{2}{\sqrt{\pi}} e^x \int_0^\infty y^{1/2} e^{-y} dy \tag{7.20}$$

The integral on the right hand side can be transformed by integrating by parts:

$$\int_0^\infty y^{1/2} e^{-y} dy = \left[-y^{1/2} e^{-y}\right]_0^\infty + \frac{1}{2} \int_0^\infty y^{-1/2} e^{-y} dy$$
$$= \frac{1}{2} \int_0^\infty y^{-1/2} e^{-y} dy$$

Making now the change of variable $Y = y^{1/2}$, and thus $dY = \frac{1}{2} y^{-1/2} dy$, we get the well-known integral:

$$\frac{1}{2} \int_0^\infty y^{-1/2} e^{-y} dy = \frac{1}{2} \int_0^\infty e^{-Y^2} dY = \frac{\sqrt{\pi}}{2}$$

Substituting in Eq. (7.20), we obtain for a non-degenerate semiconductor:

$$F_{\frac{1}{2}}(x) \approx e^x$$

and from Eqs. (7.18) and (7.20):

$$n \approx N_c \exp\left(\frac{E_F - E_C}{k_b T}\right) \tag{7.21}$$

This expression is much simpler than Eq. (7.16) and is more amenable for calculations. However, when the Fermi energy is close to or even higher than the

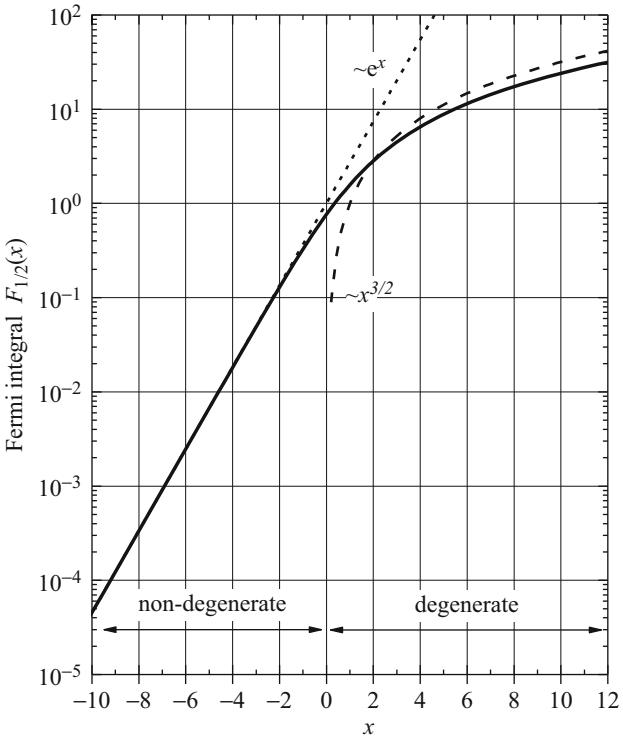


Fig. 7.2 The Fermi integral of order one half and its approximations

bottom of the conduction band, we have a so-called degenerate semiconductor, we cannot make this approximation anymore, and the Fermi-Dirac integral has to be used.

An extreme case is when $E_F - E_C >> k_bT$, corresponding to *highly degenerate* semiconductors, in which the Fermi level lies deeply inside the conduction band. Electrical properties of such semiconductors are similar to those of metals. At this condition, the Fermi-Dirac integral can be approximated as:

$$F_{\frac{1}{2}}(x) \approx x^{3/2}$$

Figure 7.2 shows the plots of the Fermi integral and the two approximations mentioned above. The exponential approximation or the 3/2 power approximation agrees very well with the Fermi integral when $x << -1$ or $x >> -1$. However, when $x \sim 0$, the Fermi-Dirac integral has to be used.

Fortunately, we are almost exclusively concerned with non-degenerate semiconductors. For example, InSb has a bandgap of 0.17 eV at 300 K, which is one of the smallest bandgaps among all the semiconductors. Assume InSb is pure and perfect (or so-called intrinsic, see Sect. 7.6), the Fermi energy is approximately

in the middle of the bandgap, $E_C - E_F \approx E_g/2$, which is about 85 meV at 300 K. Note that $k_bT = 25.9$ meV; the condition $E_C - E_F \geq 3k_bT$is satisfied. Thus the exponential form can be used. Most of the semiconductors have a larger bandgap, which means the $3k_bT$ condition is valid at room temperature.

7.4 Effective Density of States (Valence Band)

A similar derivation can be performed for the concentration or density of holes p in the valence band:

$$p = \frac{1}{V}\int_{-\infty}^{E_V} g_v(E) f_h(E) dE \tag{7.22}$$

which we obtained from Eq. (7.10) after replacing the density of states with that in the valence band and the limit of integration for an energy below the top of the valence band E_V. Moreover, the Fermi-Dirac distribution $f_e(E)$ has been replaced with (see Eq. (5.58)):

$$f_h(E) = [1 - f_e(E)] = \frac{1}{\exp\left(\frac{E_F - E}{k_bT}\right) + 1} \tag{7.23}$$

which gives the probability of the state at energy E not to be occupied by an electron and thus to be occupied by a hole.

Expanding Eq. (7.22) using Eqs. (7.8) and (7.23), we get:

$$p = \frac{1}{2\pi^2}\left(\frac{2m_h}{\hbar^2}\right)^{3/2}\int_{-\infty}^{E_V}\frac{(E_V - E)^{1/2}}{\exp\left(\frac{E_F - E}{k_bT}\right) + 1} dE \tag{7.24}$$

Using the change of variable $y = \frac{E_V - E}{k_bT}$, thus $dy = -\frac{1}{k_bT}dE$, and:

$$x = \frac{E_V - E_F}{k_bT} \tag{7.25}$$

in the previous integral and identifying it with the Fermi-Dirac integral, we obtain a relation similar to Eq. (7.17) for p:

$$p = 2\left(\frac{2\pi k_bTm_h}{h^2}\right)^{3/2} F_{\frac{1}{2}}\left(\frac{E_V - E_F}{k_bT}\right) \tag{7.26}$$

or:

$$p = N_v F_{\frac{1}{2}}\left(\frac{E_V - E_F}{k_bT}\right) \tag{7.27}$$

where:

$$N_v = 2\left(\frac{2\pi k_b T m_h}{h^2}\right)^{3/2} \tag{7.28}$$

is called the effective valence band density of states.

Example

Q Find the ratio of the heavy-hole concentration to the light-hole concentration for GaAs.

A We know that the hole concentration is related to the hole effective mass through:

$$p = 2\left(\frac{2\pi k_b T m_h}{h^2}\right)^{3/2} F_{\frac{1}{2}}\left(\frac{E_V - E_F}{k_b T}\right).$$

The Fermi-Dirac integral is the same for the heavy-hole and light-hole bands, and the only difference comes from the effective masses. Therefore, we can write: $\frac{p_{hh}}{p_{lh}} = \left(\frac{m_{hh}}{m_{lh}}\right)^{3/2}$. In GaAs, this ratio is: $\frac{p_{hh}}{p_{lh}} = \left(\frac{0.45}{0.082}\right)^{3/2} = 12.86$

Similar to what we saw in Sect. 7.3, the general expression in Eq. (7.24) can be simplified in the case of a non-degenerate semiconductor for which $E_F - E_V >> k_b T$. This situation is of most interest and is illustrated in Fig. 7.3a. It corresponds to the one where the Fermi energy is rather far from the valence band and inside the bandgap.

In this situation, the concentration of holes has a simplified expression similar to Eq. (7.18):

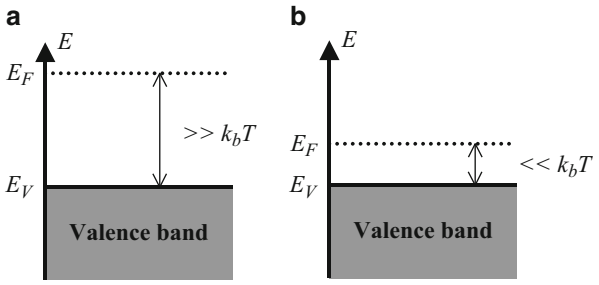


Fig. 7.3 Illustration of the position of the Fermi level with respect to the valence band (**a**) in a non-degenerate *p*-type semiconductor: The Fermi energy is far from the edge of the valence band. (**b**) In a degenerate *p*-type semiconductor, the Fermi energy is close to the edge of the valence band

$$p \approx N_{\mathrm{v}} \exp\left(\frac{E_{\mathrm{V}} - E_{\mathrm{F}}}{k_{\mathrm{b}} T}\right) \tag{7.29}$$

7.5 Mass Action Law

We saw that a non-degenerate semiconductor has its Fermi energy far away from both the bottom of the conduction band and the top of the valence band, by about a few times $k_{\mathrm{b}}T$ (25.9 meV at room temperature). This situation is more often encountered in practice than one may believe, and most of the discussion from now will therefore be in this approximation unless stated otherwise.

An important parameter is the product of n and p given in Eqs. (7.21) and (7.29) by:

$$\begin{aligned} np &= N_{\mathrm{c}} \exp\left(\frac{E_{\mathrm{F}} - E_{\mathrm{c}}}{k_{\mathrm{b}} T}\right) N_{\mathrm{v}} \exp\left(\frac{E_{\mathrm{V}} - E_{\mathrm{F}}}{k_{\mathrm{b}} T}\right) \\ &= N_{\mathrm{c}} N_{\mathrm{v}} \exp\left(\frac{E_{\mathrm{V}} - E_{\mathrm{c}}}{k_{\mathrm{b}} T}\right) \end{aligned}$$

or:

$$np = N_{\mathrm{c}} N_{\mathrm{v}} \exp\left(-\frac{E_{\mathrm{g}}}{k_{\mathrm{b}} T}\right) \tag{7.30}$$

where $E_{\mathrm{g}} = E_{\mathrm{C}} - E_{\mathrm{V}}$ is the bandgap energy of the semiconductor. This relation is very important, as it is valid for *any* value of n or p. This relation is usually called the mass action law. However, it does not hold in the degenerate semiconductor case. It is common practice to introduce the intrinsic carrier concentration, n_{i}, which is defined as:

$$n_{\mathrm{i}}^2 = np = N_{\mathrm{c}} N_{\mathrm{v}} \exp\left(-\frac{E_{\mathrm{g}}}{k_{\mathrm{b}} T}\right) \tag{7.31}$$

This parameter is a function of the semiconductor effective masses and the temperature. This concentration is qualified as "intrinsic" because for an intrinsic semiconductor, the number of electrons and holes are equal, i.e., $n = p$, and we thus have from the previous relation:

$$n = p = n_i = \sqrt{N_{\mathrm{c}} N_{\mathrm{v}}} \exp\left(-\frac{E_{\mathrm{g}}}{2k_{\mathrm{b}} T}\right) \tag{7.32}$$

Example

Q Calculate the intrinsic electron concentration for undoped GaAs at room temperature (300 K).

A For a homogeneous non-degenerate semiconductor, like undoped GaAs, the mass action law gives the intrinsic electron concentration as:

$$n_{\mathrm{i}} = \sqrt{4g_{\mathrm{d}}\left(\frac{2\pi k_{\mathrm{b}} T m_{\mathrm{e}}}{h^2}\right)^{3/2}\left(\frac{2\pi k_{\mathrm{b}} T m_{\mathrm{h}}}{h^2}\right)^{3/2}}\exp\left(-\frac{E_{\mathrm{g}}}{2k_{\mathrm{b}}T}\right)$$
$$= 2\left(\frac{2\pi k_{\mathrm{b}} T m_0}{h^2}\right)^{3/2}\sqrt{g_{\mathrm{d}}\left(\frac{m_{\mathrm{e}} m_{\mathrm{h}}}{m_0 m_0}\right)^{3/2}}\exp\left(-\frac{E_{\mathrm{g}}}{2k_{\mathrm{b}}T}\right)$$

where E_{g} is the bandgap of GaAs (1.424 eV). For GaAs, the degeneracy factor g_d is equal to 1 because the conduction band minimum is at the center of the Brillouin zone. In addition, the hole effective mass m_{h} is calculated from the heavy-hole and light-hole effective masses: $m_{\mathrm{h}}^{3/2} = m_{\mathrm{hh}}^{3/2} + m_{\mathrm{hl}}^{3/2}$. We therefore get:

$$\begin{aligned} n_i &= 2\left(\frac{2\pi\times\left(1.38066\times10^{-23}\right)\times300\times(0.91095)\times10^{-30}}{\left(6.62617\times10^{-34}\right)^2}\right)^{3/2} \\ &\times\sqrt{(0.067)^{3/2}\left(0.45^{3/2}+0.082^{3/2}\right)} \\ &\times\exp\left(-\frac{1.424\times1.60218\times10^{-19}}{2\times\left(1.38066\times10^{-23}\right)\times300}\right) \\ &= 2.06\times10^{12}\mathrm{m}^{-3} \\ &= 2.06\times10^{6}\mathrm{cm}^{-3} \end{aligned}$$

7.6 Doping: Intrinsic Versus Extrinsic Semiconductor

The energy band structures of the semiconductors that have been discussed so far corresponded to those of an intrinsic semiconductor, which is a pure and perfect semiconductor crystal. At a temperature equal to the absolute zero (0 K), the valence band of such a crystal is completely filled with electrons, and there is no electron in the conduction band. Indeed, we saw that the Fermi energy of a semiconductor lies within a forbidden energy gap (Sub-sect. 5.2.7). Since the Fermi-Dirac distribution function has an exact step shape at $T = 0$ K (Fig. 5.12), there is no electron with an energy $E > E_{\mathrm{F}}$, including the conduction band, and all the electrons are located at an energy $E < E_{\mathrm{F}}$.

This phenomenon directly results from the fact that the outer shell of each constituent atom of a semiconductor is fully filled with four electrons. Counting the number in the four shared bonds then gives a total of eight electrons. For example, in the case of a silicon crystal, illustrated in Fig. 7.4, each Si atom is bonded to four neighboring Si atoms. A Si atom originally has four electrons in its outer shell (it is in the column IV of the periodic table), each of which is shared with

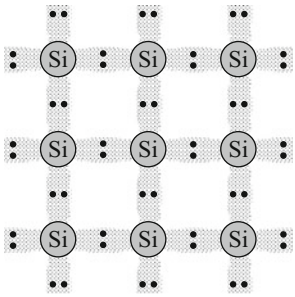


Fig. 7.4 Schematic of a Si semiconductor crystal showing the distribution of electrons in the outer shell of each Si atom. Each Si atom has eight electrons in this shell: four from its own outer shell and one from each of the four nearest Si atoms to which it is covalently bonded to

one different neighboring atom. Every Si atom has therefore a total of eight electrons: Its original four electrons and one electron from each of the four neighboring Si atoms.

We thus see that all the outer shell electrons are shared into bonds, and thus there is no extra free electron which can move. Moreover, all the outer shell "spots" are filled with electrons; therefore there is no room for an electron to move to if displaced by a field. As a result, the electrical conductivity of a pure semiconductor is "low" (only excited states can conduct). This is why a pure semiconductor is an insulator at the absolute zero temperature.

In order to either increase the number of free electrons or increase the number of "spots" (empty energy levels) where a potential electron can move into, we need to replace some of the Si atoms with other elements, called dopants, which are not isoelectronic to it, i.e., not with the same number of outer shell electrons. This process is called doping, which results in an extrinsic semiconductor. A dopant is thus an impurity added to the semiconductor crystal. Because the dopant replaces or substitutes a Si atom, it is called a substitutional dopant. The concentration of such dopants typically introduced in a semiconductor is in the range of 10^{15}–10^{19} cm^{-3}, which is low in comparison with the concentration of atoms in a crystal (typically ~10^{22} cm^{-3}). There are two types of doping, *n*-type doping and *p*-type doping, depending on the nature of the dopant introduced. Such a dopant can be introduced intentionally or unintentionally during the synthesis of the semiconductor crystal.

The *n*-type doping is achieved by replacing a Si atom with an atom with *more* electrons in the outer shell. This can be achieved, for example, by using phosphorus (P), an element from the column V of the periodic table, which has five electrons in its outer shell. The result is shown in Fig. 7.5.

As we can see, four of the electrons in the outer shell of the P atom are involved in covalent bonds with its four neighboring Si atoms. The fifth electron is therefore free to move in space. The P atom is therefore called a donor in silicon because it can give away an electron which can in turn participate in electrical conductivity phenomena. Once an electron is given away, the phosphor atom becomes a positively charged ion and is then called an ionized donor. This ionization process is generally achieved through thermal excitation of an electron from the outer shell of the donor atom.

Because the dopant creates a perturbation to the periodicity of the crystal lattice, it gives rise to additional energy levels in the bandgap. When the dopant concentration

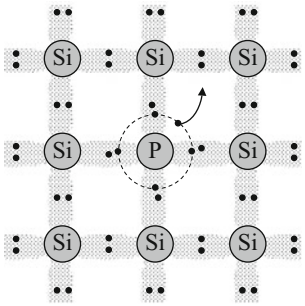


Fig. 7.5 Schematic of a Si semiconductor crystal with one Si atom replaced by a P atom to achieve *n*-type doping. The dotted circle symbolizes the outer shell of the P atom which contains five electrons. Because the fifth electron does not contribute to the bonding, it can be free (ionized) to move inside the crystal. P is thus called a donor

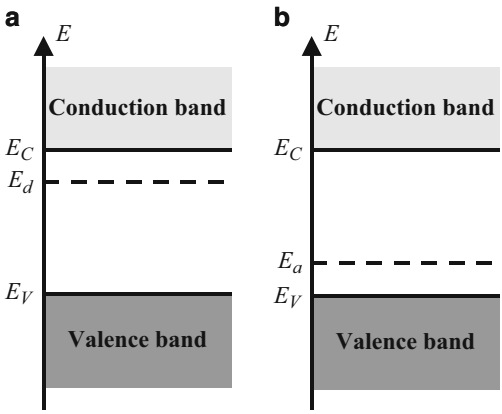


Fig. 7.6 Schematic of the energy levels introduced by (**a**) a donor or (**b**) an acceptor dopant in a semiconductor crystal. The energy level of a donor is closer to the edge of the conduction band, whereas that of an acceptor is closer to the edge of the valence band

is low in comparison with the density of crystal atoms, the dopant energy level can be considered as isolated, i.e., there is no energy band associated with it. We can then talk about a donor energy level E_d, as shown in Fig. 7.6a. Moreover, because the extra electron around the P atom is easily ionized, it has a small binding energy, with respect to the conduction band. The energy of the donor electron E_d is closer to the conduction band than the valence band. The ionization energy of the dopant is the difference $E_C - E_d$.

The other type of doping, *p*-type doping, is achieved by replacing a Si atom with an atom with *fewer* electrons in the outer shell. This can be achieved, for example, by using gallium (Ga), an element from the column III of the periodic table, which has three electrons in its outer shell. The result is shown in Fig. 7.7.

As we can see, all three electrons in the outer shell of the Ga atom are involved in covalent bonds with three of its four neighboring Si atoms. There thus remains an open location that can be filled with an electron. The Ga atom is therefore called an acceptor in silicon because it can "accept" or "capture" an extra electron from a

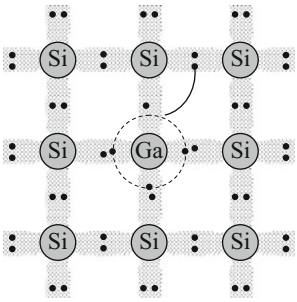


Fig. 7.7 Schematic of a Si semiconductor crystal with one Si atom replaced by a Ga atom to achieve p-type doping. The dotted circle symbolizes the outer shell of the Ga atom which contains three electrons. The Ga atom can accept one more electron from a neighboring bond. Ga is thus called an acceptor

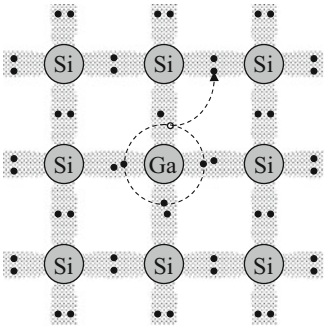


Fig. 7.8 Schematic showing the movement of a hole in a Si semiconductor crystal doped *p*-type using a Ga atom. The hole is represented by an open circle. When the Ga atom accepts an electron, the process can be equivalently viewed as the Ga atom releasing a hole inside the crystal

neighboring covalent bond, thus leaving a new available location for electron capture. Once an electron is captured, the gallium atom becomes a negatively charged ion and is then called an ionized acceptor. This movement of electrons is involved in electrical conductivity phenomena. Remembering the concept of holes discussed in Sub-sect. 5.3.3, we can see that this electron movement is equivalent to the movement of a hole in the opposite direction, as illustrated in Fig. 7.8. The Ga atom, an acceptor (of electrons) in silicon, can then be also considered as a donor of holes.

Here again, the *p*-type dopant is a perturbation of the periodicity of the crystal lattice and leads to additional localized energy levels (i.e., not bands) in the bandgap at E_a, which is called acceptor energy level, as shown Fig. 7.6b. Because the Ga atom easily captures an electron, E_a is closer to the valence band than the conduction band. The ionization energy of the *p*-type dopant is the difference $E_a - E_V$.

A semiconductor may contain donors (with a concentration N_D) and acceptors (with a concentration N_A) at the same time. We then talk about compensation and say that the semiconductor is compensated. The overall behavior of this semiconductor depends on the relative difference between N_D and N_A. In either case of *n*-type

Table 7.1 Dopants and ionization energies for (a) Si, (b) Ge, (c) GaAs (Sze 1981; Wolfe et al. 1989), and (d) InP (http://www.ioffe.ru/SVA/NSM/Semicond/InP/index.html)

(a) Si		
Impurity	Type	Ionization energy (meV)
P	Donor	45.31
As	Donor	53.51
Sb	Donor	42.51
B	Acceptor	45
Al	Acceptor	57
Ga	Acceptor	65
(b) Ge		
Impurity	Type	Ionization energy (meV)
P	Donor	12.76
As	Donor	14.04
Sb	Donor	10.19
B	Acceptor	10.47
(c) GaAs		
Impurity	Type	Ionization energy (meV)
Si	Donor	5.854
Ge	Donor	5.908
S	Donor	5.89
Be	Acceptor	30
Mg	Acceptor	30
Zn	Acceptor	31.4
C	Acceptor	26.7
(d) InP		
Impurity	Type	Ionization energy (meV)
Si	Donor	5.7
S	Donor	5.7
Sn	Donor	5.7
Be	Acceptor	30
Mg	Acceptor	30
Zn	Acceptor	35

and/or *p*-type doping, the mass action law expressed in Eq. (7.30) remains valid as long as we have a non-degenerate semiconductor.

Table 7.1 lists the most common dopants with their ionization energies for the following semiconductors: Si, Ge, GaAs, and InP.

7.7 Charge Neutrality

A semiconductor crystal, be it intrinsic or extrinsic, must be electrically neutral at a macroscopic scale. Indeed, even if dopants are introduced, they are electrically neutral, and therefore the semiconductor crystal remains globally neutral too. As the dopants get ionized, they create mobile electrons and holes in the crystal. But,

there is no persistent accumulation of electrical charges. Even in a compensated semiconductor, overall charge neutrality remains.

Before mathematically expressing the electrical neutrality condition, we must first count all the electrical charges present in the crystal. The negative charges include the electrons in the conduction band, with a concentration n, and the ionized acceptors with a concentration N_A^-. The positive charges include the holes in the valence band, with a concentration p, and the ionized donors with a concentration N_D^+. The charge neutrality relation can then be written as:

$$n + N_A^- = p + N_D^+ \tag{7.33}$$

For a given semiconductor crystal, the concentrations n and p solely depend on the Fermi energy E_F through Eqs. (7.21) and (7.29) in the non-degenerate case or Eqs. (7.18) and (7.27) in the general case. The concentrations of ionized donors N_D^+ and acceptors N_A^- depend also on the Fermi energy for a given dopant nature, the temperature T, and total concentration as follows:

$$\frac{N_D^+}{N_D} = \frac{1}{2\exp\left(\frac{E_F - E_d}{k_b T}\right) + 1} \tag{7.34}$$

$$\frac{N_A^-}{N_A} = \frac{1}{4\exp\left(\frac{E_a - E_F}{k_b T}\right) + 1} \tag{7.35}$$

where E_F is the Fermi energy, E_d and E_a are the donor and acceptor energy levels in the bandgap, respectively, and N_D and N_A are the total donor and acceptor concentrations, respectively. The factor 2 in Eq. (7.34) arises because the donor atom can in practice only be singly occupied by an electron (electron-electron repulsion will prevent double occupation), and the factor 4 in Eq. (7.35) arises for the same reason and the fact that there are two degenerate subbands in the valence band at the center of the Brillouin zone: the heavy-hole band and the light-hole band (Sub-sect. 5.4.3). Similar to the Fermi-Dirac distribution, Eqs. (7.34) and (7.35) are derived from statistical physics.

The charge neutrality equation is a very important property because it gives an implicit equation which can be used to determine the Fermi energy. Once the Fermi energy is determined, the concentration of electrons in the conduction band and that of holes in the valence band can be readily calculated through Eqs. (7.21) and (7.29) in the non-degenerate case or Eqs. (7.18) and (7.27) in the general case.

7.8 Fermi Energy as a Function of Temperature

An example of such calculation is given here, first for an intrinsic and then for an n-type extrinsic and non-degenerate semiconductor.

In the intrinsic case, we assume there is no dopant, i.e., the total concentration of dopant is $N_D = N_A = 0$. Substituting in Eq. (7.33), we therefore obtain Eq. (7.32) again. Now, by identifying n in Eqs. (7.21) and (7.32), we can write an expression for the Fermi energy:

$$n = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2k_bT}\right) = N_c \exp\left(\frac{E_F - E_C}{k_bT}\right)$$

which becomes, knowing that $E_g = E_C - E_V$:

$$\exp\left(\frac{E_F}{k_bT}\right) = \sqrt{\frac{N_v}{N_c}}\exp\left(\frac{E_C + E_V}{2k_bT}\right) \tag{7.36}$$

After taking the logarithm of this relation:

$$\frac{E_F}{k_bT} = \frac{E_C + E_V}{2k_bT} + \ln\left(\sqrt{\frac{N_v}{N_c}}\right)$$

or:

$$E_F = \frac{E_C + E_V}{2} + \frac{1}{2}k_bT \ln\left(\frac{N_v}{N_c}\right) \tag{7.37}$$

This equation shows that the Fermi energy in an intrinsic semiconductor lies near the middle of the bandgap and is offset by an amount that varies with temperature. At the absolute zero temperature, the Fermi energy is exactly at the middle of the bandgap.

Example

Q Determine how far the Fermi energy is from the middle of the bandgap of GaAs at 296 K.

A The Fermi energy is given the expression: $E_F = \frac{E_C+E_V}{2} + \frac{1}{2}k_bT \ln\left(\frac{N_v}{N_c}\right)$. The energy difference between the Fermi energy and the middle of the bandgap is therefore given by the logarithm function, $E_F - \frac{E_C+E_V}{2} = \frac{1}{2}k_bT \ln\left(\frac{N_v}{N_c}\right)$, which is given by the ratio: $\frac{1}{2}k_bT \ln\left(\frac{1}{g_d}\left(\frac{m_h}{m_e}\right)^{3/2}\right)$. In GaAs, the degeneracy factor g_d is equal to 1 because the conduction band minimum is at the center of the Brillouin zone. In addition, the hole effective mass m_h is calculated from the heavy-hole and light-hole effective masses: $m_h^{3/2} = m_{hh}^{3/2} + m_{lh}^{3/2}$. This leads to:

$$E_F - \frac{E_C + E_V}{2} = \frac{1}{2} k_b T \ln\left(\frac{m_{hh}^{3/2} + m_{lh}^{3/2}}{m_e^{3/2}}\right) = \frac{1}{2}$$
$$\times 1.38066 \times 10^{-23} \times 296 \times \ln\left(0.45^{3/2} + \frac{0.082^{3/2}}{0.067^{3/2}}\right)$$
$$= 6.00 \times 10^{-21} \text{J} = 37.4 \text{meV}$$

For an extrinsic semiconductor, an expression similar to Eq. (7.37) cannot be easily obtained, because one needs to estimate the concentrations of ionized donors (N_D^+) or acceptors (N_A^-) as a function of the total concentrations, which is beyond the scope of this textbook. Nevertheless, the following discussion will enable us to qualitatively understand the variation of the Fermi energy as a function of temperature.

We know that, at the absolute zero temperature ($T = 0$ K), the Fermi energy E_F is such that all the electrons have an energy below E_F and no electron has an energy higher than E_F.

Therefore, in an *n*-type doped semiconductor at $T = 0$ K, the Fermi energy is located between E_C and E_d, as illustrated in Fig. 7.9a, which means that the Fermi energy is much closer to the bottom of the conduction band than in the case of an intrinsic semiconductor. This proximity has the very important consequence that the concentration of electrons in the conduction band is much larger than for an intrinsic semiconductor, as a result of the shape of the Fermi-Dirac distribution shown in Fig. 5.12, when the temperature is raised. These electrons can easily participate in electrical conduction phenomena.

By contrast, in a *p*-type doped semiconductor at $T = 0$ K, the Fermi energy E_F is located between E_V and E_a, as illustrated in Fig. 7.9b, which means that the Fermi energy is much closer to the top of the valence band than in the case of an intrinsic

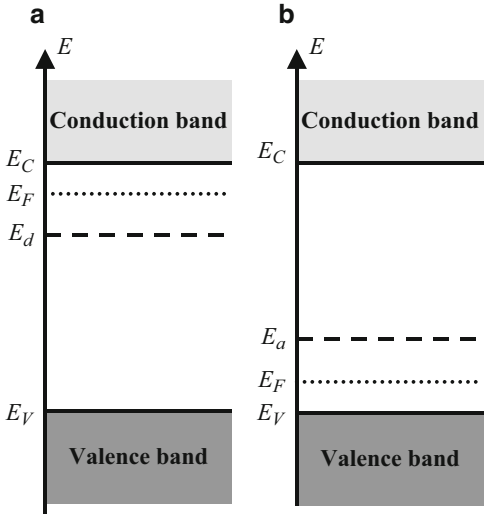


Fig. 7.9 Position of the Fermi energy at $T = 0$ K in (**a**) an n-type semiconductor is located between the donor energy level and the bottom of the conduction band, and (**b**) in a *p*-type semiconductor, it is located between the acceptor energy level and the top of the valence band

semiconductor. This proximity also has the very important consequence that the concentration of holes in the valence band is much larger than for an intrinsic semiconductor, as a result of the shape of the Fermi-Dirac distribution shown in Fig. 5.12, at room temperature. And these holes can easily participate in electrical conduction phenomena.

For very high temperatures, all the donors or acceptors are ionized, and we have $N_D{}^+ = N_D$ or $N_A{}^+ = N_A$. Thus, the contribution from dopants to the charged carriers is limited, which is typically to a maximum of 10^{19} cm^{-3}. At the same time, the intrinsic contribution to the concentrations of electrons and holes, given by Eq. (7.32), is such that (take $T \to \infty$):

$$n = p = n_i = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2k_b T}\right) \approx \sqrt{N_c N_v} \tag{7.38}$$

Moreover, from Eqs. (7.19) and (7.28), we saw that the effective density of states N_c and N_v both increase as $T^{3/2}$. Therefore, the intrinsic contribution to n and p also increases as $T^{3/2}$, i.e., is not limited when the temperature increases, unlike the contribution from dopants. The charge neutrality relation in Eq. (7.33) then becomes:

$$n \approx p \tag{7.39}$$

This means that at very high temperatures, the charge carriers in an extrinsic semiconductor behave as in an intrinsic semiconductor. This also means that the Fermi energy tends to the expression given in Eq. (7.37). From these qualitative arguments, we can schematically illustrate the evolution of the Fermi energy as a function of temperature in Fig. 7.10 for an *n*-type and a *p*-type semiconductor.

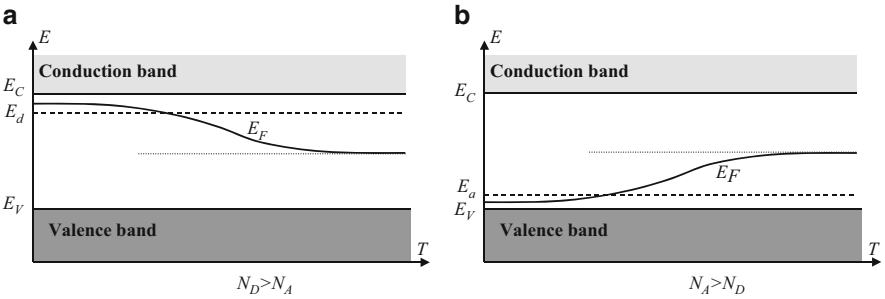


Fig. 7.10 Evolution of the Fermi energy as a function of temperature in a (**a**) n-type or (**b**) p-type semiconductor crystal. As the temperature is raised, the position of the Fermi energy shifts from its position in Fig. 7.9 to the position for an intrinsic semiconductor

7.9 Carrier Concentration in an *n*-Type Semiconductor

Before concluding this section on the electrical charge distribution at equilibrium, let us consider the example of a non-degenerate, *n*-type doped semiconductor. Here again, we will not go in a detailed numerical analysis but will provide the main qualitative results. The total dopant concentration will be denoted N_D. Assuming there is no acceptor ($N_A = 0$), the charge neutrality relation in Eq. (7.33) is now:

$$n = p + N_D^+ \tag{7.40}$$

Several levels of approximations, corresponding to several temperature regimes, can be considered to further simplify this expression. But before continuing the discussion, we should point out that holes in this semiconductor can only originate from the intrinsic contribution, not an extrinsic source such as a dopant (we chose $N_A = 0$).

The first regime corresponds to high temperatures. As discussed in the previous subsection, all the donors are ionized ($\left(N_D^+ = N_D\right)$). However, the concentrations of electrons n and holes p are much higher than the total concentration of donors (n, $p >> N_D$), and they therefore obey the expressions derived for the intrinsic case, i.e.:

$$n \approx p \approx n_i = \sqrt{N_c N_v}\exp\left(-\frac{E_g}{2k_b T}\right) \tag{7.41}$$

As the temperature is lowered, while the donors remain ionized $\left(N_D^+ = N_D\right)$, the intrinsic contribution to the concentrations of electrons and holes diminishes. Below a certain temperature, these contributions become negligible in comparison to N_D^+ or N_D. In this second temperature regime, p can be neglected ($p \ll N_D^+$) because the only contribution to p is the intrinsic contribution. Therefore, Eq. (7.40) becomes:

$$n \approx N_D \tag{7.42}$$

This is the most interesting characteristic of an extrinsic semiconductor. Indeed, if the concentration of donors can be intentionally controlled in the crystal during the synthesis, the concentration of electrons in the conduction band is precisely determined.

Specifically, the temperature at which the carrier concentration from thermal generation becomes equal to the background carrier concentration is called the intrinsic temperature T_i. Below T_i the carrier concentration is relatively temperature independent. Above T_i it increases exponentially with temperature.

As the temperature is further lowered, we reach a third regime where all the donors are not ionized anymore ($N_D^+ < N_D$). At the same time, we still have $p \ll N_D^+$. In this case, Eq. (7.40) becomes:

$$n \approx N_D^+ \tag{7.43}$$

At low temperatures, the Fermi energy E_F lies between the bottom of the conduction band E_C and the donor level E_d. Therefore, $E_F - E_d > 0$ and the expression for N_D^+ in Eq. (7.34) can be simplified to become:

$$\frac{N_D^+}{N_D} = \frac{1}{2\exp\left(\frac{E_F - E_d}{k_b T}\right) + 1} \approx \frac{1}{2\exp\left(\frac{E_F - E_d}{k_b T}\right)}$$

or:

$$N_D^+ \approx \frac{N_D}{2}\exp\left(-\frac{E_F - E_d}{k_b T}\right) \tag{7.44}$$

Let us now calculate the product nN_D^+. On the one hand, it is equal to n^2 from Eq. (7.43). On the other hand, it is equal to:

$$N_c\exp\left(\frac{E_F - E_C}{k_b T}\right)\frac{N_D}{2}\exp\left(-\frac{E_F - E_d}{k_b T}\right) \tag{7.45}$$

after using Eqs. (7.21) and (7.44). We then obtain:

$$n^2 \approx \frac{N_c N_D}{2}\exp\left(-\frac{E_C - E_d}{k_b T}\right) \tag{7.46}$$

which yields:

$$n \approx \sqrt{\frac{N_c N_D}{2}}\exp\left(-\frac{E_C - E_d}{2k_b T}\right) \tag{7.47}$$

The three expressions of n in Eqs. (7.41), (7.42), and (7.47) provide good approximations of the concentration of electrons in the conduction band as a function of temperature. It is customary to plot this concentration in a logarithmic scale for n and as a function of inverse temperature (i.e., $\frac{1}{k_b T}$), so that the slopes of the curve can be directly correlated to the bandgap energy E_g in Eq. (7.47) and the ionization energy $E_C - E_d$ in Eq. (7.47)). This is very simply shown in the schematic diagram in Fig. 7.11. Here, the temperature dependence of N_c ($T^{3/2}$) from Eq. (7.19)) has been neglected in comparison to the temperature dependence of the exponential terms.

In the case of a *p*-type semiconductor, with an acceptor concentration N_A, the following hole concentrations for the various regimes discussed previously can be determined.

In the first regime, at high temperatures, the concentrations of holes p and electrons n are much higher than the total concentration of acceptors ($n, p >> N_A$) and thus follow their expressions for the intrinsic case, as in Eq. (7.41):

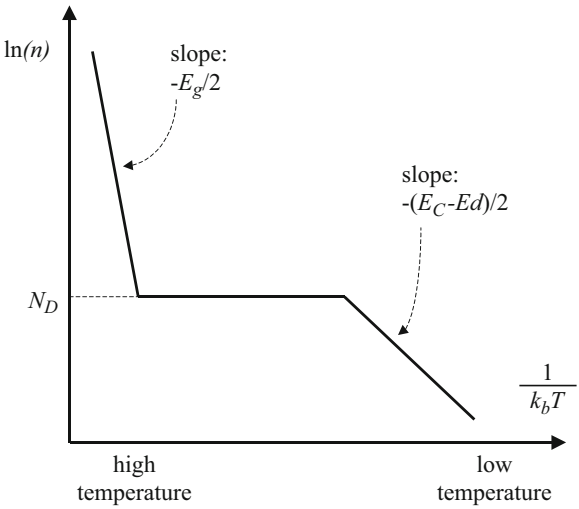


Fig. 7.11 Simple schematic diagram of the dependence of the electron concentration in the conduction band as a function of temperature in a typical *n*-type semiconductor crystal. At low temperatures, the carrier concentration follows a relation dependent on the donor energy inside the bandgap. At moderate temperatures, the electron concentration is nearly constant equal to the donor concentration. At high temperatures, the carrier concentration approaches that of an intrinsic semiconductor

$$n \approx p \approx n_{\mathrm{i}} = \sqrt{N_{\mathrm{c}} N_{\mathrm{v}}} \exp\left(-\frac{E_{\mathrm{g}}}{2k_{\mathrm{b}}T}\right) \tag{7.48}$$

In the second regime, Eq. (7. 42) can be transformed for a *p*-type semiconductor into:

$$p \approx N_{\mathrm{A}} \tag{7.49}$$

In the third regime, as the temperature is further lowered, Eq. (7.43) can also be transformed for a *p*-type semiconductor into:

$$p \approx N_{\mathrm{A}}^{-} \tag{7.50}$$

From Eq. (7.35), using the same derivation as between Eqs. (7.44) and (7.47), we get:

$$p \approx \sqrt{\frac{N_{\mathrm{v}} N_{\mathrm{D}}}{4}} \exp\left(-\frac{E_{\mathrm{a}} - E_{\mathrm{V}}}{2k_{\mathrm{b}}T}\right) \tag{7.51}$$

7.10 Summary

In this chapter, we have first described the equilibrium properties of charge carriers in a semiconductor. We introduced the concepts of effective density of states, mass action law, and intrinsic and extrinsic semiconductor. The *n*-type and *p*-type doping of semiconductors has been discussed, taking into account the charge neutrality of the solid. We also discussed the importance of the Fermi energy.

Problems

1. Calculate the conduction band effective density of states for Si, Ge, and GaAs at 300 K. Plot it in logarithmic scale as a function of the logarithm of the temperature.
2. Calculate the valence band effective density of states for Si, Ge, and GaAs at 300 K. Plot it as a function of temperature, in logarithmic scale. We know that the valence band is degenerate at the center of the Brillouin zone as there is a heavy-hole band (with effective mass m_{hh}) and a light-hole band (with effective mass m_{hh}). The effective mass to be used in Eq. (7.28) is then:

$$m_{\mathrm{h}} = \left(m_{\mathrm{hh}}^{3/2} + m_{\mathrm{lh}}^{3/2} \right)^{2/3}.$$

3. Find the energies at which the distribution of electrons in the conduction band and the distribution of holes in the valence band have maxima, if distributions are governed by Maxwell-Boltzmann statistics.
4. Estimate relative errors in the calculation of free carrier concentration when the Maxwell-Boltzmann statistics is applied for semiconductors with Fermi energy within the energy gap, if the Fermi level is $3k_{\mathrm{b}}T$, $2k_{\mathrm{b}}T$, and $k_{\mathrm{b}}T$ away from the bandgap edge or if it coincides with the edge. Use the given table of the exact value of the Fermi integral ($F_{1/2}$) for the comparison.
5. Calculate the intrinsic carrier concentrations for Si, Ge, GaAs, and GaN at 300 K, in the non-degenerate case. Plot their evolution as a function of temperature, in logarithmic scale.
6. From the periodic table, give examples of *n*-type and *p*-type dopants for Ge and GaAs. Is silicon an *n*-type or a *p*-type dopant in GaAs? Interpret.
7. As we know P is an *n*-type dopant for Si and Ge. Nitrogen is in the same column as P in the periodic table. Will N be a good dopant? Why?
8. Give an expression for the charge neutrality relation when double acceptors are present with a concentration N_{AA}. Double acceptors accept one or two electrons. Use the same notations as those in Sect. 7.3.
9. Plot the evolution of the Fermi energy as a function of temperature in intrinsic GaAs.
10. Consider a *p*-type doped GaAs semiconductor at 300 K with an experimentally measured hole concentration of 1.5×10^{17} cm^{-3}. The *p*-type dopant has an energy level such that $\Delta E_{\mathrm{a}} = E_{\mathrm{a}} - E_{\mathrm{V}} = 125$ meV. Assuming there is no donor, determine the proportion of ionized acceptors. Determine the total concentration of acceptors.
11. Consider an *n*-type doped GaAs semiconductor at 300 K with an experimentally measured electron concentration of 3×10^{17} cm^{-3}. The *n*-type dopant has an energy level such that $\Delta E_{\mathrm{d}} = E_{\mathrm{C}} - E_{\mathrm{d}} = 25$ meV. Assuming there is no acceptor, determine the proportion of ionized donors. Determine the total concentration of donors.

12. Derive expressions for concentrations of free carriers in a semiconductor doped with both, donor and acceptor impurities. Determine the conductivity type and calculate the concentrations of carriers in silicon at $T = 300$ K, if it is doped with:
    (a) $N_A = 10^{16}$ cm^{-3} >> N_D.
    (b) $N_D = 10^{16}$ cm^{-3} >> N_A.
    (c) $N_D = N_A = 10^{16}$ cm^{-3}.
    Assume that all impurities are ionized and $n_i = 1.38 \times 10^{10}$ cm^{-3} at 300 K.
13. Calculate the concentration of acceptor impurities in silicon, and determine the type of semiconductor, if at $T = 300$ K the concentration of electrons is 5×10^{11} cm^{-3} and the concentration of donor impurities is 10^{15} cm^{-3}. Assume $n_i = 1.38 \times 10^{10}$ cm^{-3} at 300 K.
14. Calculate concentrations of carriers in silicon doped by acceptors $N_A = 10^{14}$ cm^{-3} at:
    (a) 27 °C
    (b) 175 °C

References

Sze SM (1981) *Physics of Semiconductor Devices*. Wiley, New York
Wolfe CM, Holonyak N Jr, Stillman GE (1989) Physical properties of semiconductors. Prentice-Hall, Englewood Cliffs

Further Reading

Anselm A (1981) Introduction to semiconductor theory. Prentice-Hall, Englewood Cliffs
Ashcroft NW, Mermin ND (1976) *Solid State Physics*. Holt, Rinehart and Winston, New York
Cohen MM (1972) Introduction to the quantum theory of semiconductors. Gordon and Breach, New York
Ferry DK (1991) Semiconductors. Macmillan, New York
Hummel RE (1986) Electronic properties of materials. Springer, New York
Pierret RF (1989) Advanced semiconductor fundamentals. Addison-Wesley, Reading
Sapoval B, Hermann C (1995) Physics of Semiconductors. Springer, New York
Streetman BG (1990) Solid state electronic devices. Prentice-Hall, Englewood Cliffs
Wang S (1989) Fundamentals of semiconductor theory and device physics. Prentice-Hall, Englewood Cliffs

Non-equilibrium Electrical Properties of Semiconductors

8

8.1 Introduction

In the previous chapter, we established the basic relations and formalism for the distribution of electrons in the conduction band and holes in the valence band at thermal equilibrium.

Although the equilibrium state for electrons and holes in a semiconductor is the result of interactions between carriers or between carriers and phonons, it does not depend on the way this state is reached. The knowledge of the equilibrium properties is therefore not sufficient, and this is all the more true since semiconductor devices usually work under non-equilibrium conditions. In this chapter, we will thus discuss the dynamics of electrons and holes, including electrical conductivity, Hall effect, diffusion, as well as recombination mechanisms.

8.2 Electrical Conductivity

8.2.1 Ohm's Law in Solids

Because electrons and holes are charged particles, they can move in an orderly manner in a semiconductor under the influence of an electric field, for example. This motion generates an electrical current, called drift current, which is at the origin of the electrical conductivity phenomenon of certain solids. The magnitude of this current determines whether a solid is a "good" or a "bad" conductor and is directly related to the density of mobile electrical charge carriers in the solid. In this section, we will try to model the electrical conductivity in solids starting from the Drude model, which is a general model and is valid for any solid which contains mobile charge carriers. This model is based on the kinetic theory of gases which was briefly mentioned in Sect. 6.11.

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_8

In this model, an electron from the gas of electrons is considered as (i) a free moving particle in space with a momentum and an energy, (ii) which is subject to instantaneous collision events (e.g., with other particles such as electrons or atom cores or with irregularities in the crystal), (iii) the probability for a collision to occur during an interval of time dt is proportional to dt, (iv) and the particles reach their thermal equilibrium only through these collisions (see the Monte Carlo method in Appendix A.8).

Let us start by conceptually considering an electron with an electrical charge $-q$ in a uniform electric field strength $\vec{E}$. The force exerted on this electron is constant and equal to $-q\,\vec{E}$ ($q > 0$). Newton's action mass law is such that:

$$m\frac{d\,\vec{v}}{dt} = \vec{F} = -q\,\vec{E} \tag{8.1}$$

where $\vec{v}$ is the velocity of the electron and m is its mass (in a semiconductor $m = m_e$, the effective mass). This relation means that the acceleration of the electron is constant and therefore that its velocity increases linearly with time. In practice the velocity does not increase indefinitely, because collisions, which change the energy and or scatter the momentum, prevent the electron velocity from reaching extremely high values.

The current density vector $\vec{J}$ is a vector which is parallel to the flow of charge and whose magnitude is equal to the amount of electrical charge (in Coulomb) that passes per unit time through a unit area surface perpendicular to the flow of charges, as shown in Fig. 8.1a. The current density is expressed in units of $\mathrm{A{\cdot}cm^{-2}}$.

The current density can be determined by calculating the number of electrons which will traverse the surface S, during a time interval dt. Such electrons are in fact located in the volume defined between the surfaces denoted by S and S' in Fig. 8.1b. This volume is equal to $A\left|\vec{v}\right|dt$, where A is the area of the surface S.

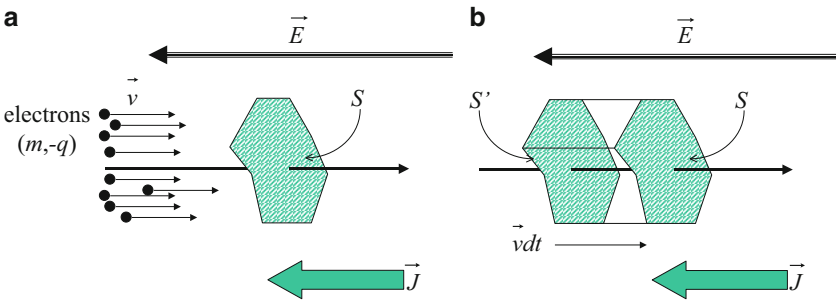


Fig. 8.1 Schematic diagrams showing (**a**) the flow of electrons and current density vector in a uniform electric field, (**b**) the displacement of the surface area A after a time dt at a velocity equal to that of the flowing electrons

Assuming that there is a concentration n of electrons in this region of space and that all of them have a velocity $\vec{v}$, the total amount of electrical charge traversing the surface S with area A, during a time interval dt, is:

$$nqA\left|\vec{v}\right|dt \tag{8.2}$$

The magnitude of the current density is thus the expression in Eq. (8.2) divided by the area and the time interval. Because the current density vector is parallel and in opposite direction to the flow of electrons, we obtain:

$$\vec{J} = -nq\,\vec{v} \tag{8.3}$$

In reality, the electrons are subject to collisions and do not all have the same velocity $\vec{v}$ individually, but they can be considered to have the same averaged velocity, and the expression in Eq. (8.3) remains valid by considering that $\vec{v}$ is the average velocity of the electron gas as a whole. Indeed, if there were no electric field, because collisions are a statistical process, the electrons are as likely to move in one direction in space as in another after a collision. The average velocity vector of the electron gas is thus zero, and there would be no electrical current, as expected (see the Monte Carlo method in Appendix A.8).

In order to calculate the average velocity of the electron gas that results from the electric field, we have to introduce, as was done earlier in Chap. 6, a characteristic time called electron relaxation time τ, which is the average duration between two consecutive collisions or scattering events. Such durations typically range on the order of 10^{-12}–10^{-14} s for electrons in metals. The probability of a collision to occur is in fact proportional to $\frac{1}{\tau}$. The average velocity is then called drift velocity and is denoted $\vec{v}^{\text{drift}}$. This quantity can be estimated by integrating Eq. (8.1) over time from $t = 0$ and $t = \tau$:

$$m\,\vec{v}^{\text{drift}} = -q\tau\,\vec{E} \text{ or } \vec{v}^{\text{drift}} = -\frac{q\tau}{m}\,\vec{E} \tag{8.4}$$

We see that the drift velocity is proportional to the electric field strength and this proportionality factor is called the mobility of electrons in the solid:

$$\begin{cases} \vec{v}^{\text{drift}} = -\mu\,\vec{E} \\ \mu = \dfrac{q\tau}{m} \end{cases} \tag{8.5}$$

This quantity is expressed in units of $\text{cm}^2\cdot\text{V}^{-1}\cdot\text{s}^{-1}$, and it represents the velocity that an electron gains per unit electric field strength (velocity ($\text{cm}\cdot\text{s}^{-1}$) divided by electric field strength ($\text{V}\cdot\text{cm}^{-1}$)). This parameter is not used often in metals but will be most useful to characterize semiconductors. The drift current density, which results from the drift of electrons in the electric field, can then be written using Eq. (8.3):

$$\vec{v}^{\text{drift}} = -nq\ \vec{v}^{\text{drift}} = nq\mu\ \vec{E} = \frac{nq^2\tau}{m}\vec{E} \tag{8.6}$$

A "drift" superscript has been added to emphasize that this is the drift current density. Here again, we see that the current density is proportional to the electric field strength. This proportionality factor is called the conductivity, is denoted σ, and is expressed in units of $\text{S}\cdot\text{cm}^{-1}$ (Siemens per cm) or inverse ($\Omega\cdot$cm):

$$\begin{cases} \vec{J}^{\text{drift}} = \sigma\vec{E} \\ \sigma = nq\mu = \dfrac{nq^2\tau}{m} \end{cases} \tag{8.7}$$

It is also a common practice to consider the inverse of the conductivity which is called the resistivity of the material:

$$\rho = \frac{1}{\sigma} = \frac{1}{nq\mu} \tag{8.8}$$

The linear relation in Eq. (8.7) is called Ohm's law. In strong electric fields, deviations from this linear dependence may occur, but one can keep the general expression for the current density in Eq. (8.7) by considering a field-dependent conductivity σ. In this case, the relation is called the generalized Ohm's law.

Example

Q Estimate the electron mobility in Cu.

A The charge carriers in Cu are electrons, and their mobility μ is related to the resistivity ρ of Cu through $\mu = \frac{1}{nq\rho}$, where n is the electron concentration participating in the conduction. Since there are two electrons in the valence shell of copper, this concentration can be determined by the concentration of Cu atoms or the density of Cu ($d = 8.92\ \text{g}\cdot\text{cm}^{-3}$) through $n = 2 \times \frac{d}{m_{\text{Cu}}}$, where m_{Cu} is the mass of a Cu atom. Assuming the resistivity of Cu is about $\rho = 1.7 \times 10^{-6}\ \Omega\cdot$cm, we get the mobility:

$$\begin{aligned} \mu &= \frac{m_{\text{Cu}}}{2dq\rho} \\ &= \frac{63.55 \times 1.67264 \times 10^{-27}}{2 \times \left(8.92 \times 10^{-3}\right) \times \left(1.60218 \times 10^{-19}\right) \times \left(1.7 \times 10^{-6}\right)} \\ &= 21.9\,\text{cm}^2/\text{Vs} \end{aligned}$$

For many, Ohm's law is more commonly recognized through the relation "$I = \frac{V}{R}$," where I is the current, V the voltage, and R the resistance of an electrical component. Indeed, let us consider a parallelepiped-shaped solid, as depicted in Fig. 8.2. We assume the electric field in the solid is uniform and that the electrical current flows perpendicularly to a side of the parallelepiped with surface area WH, as shown in Fig. 8.2.

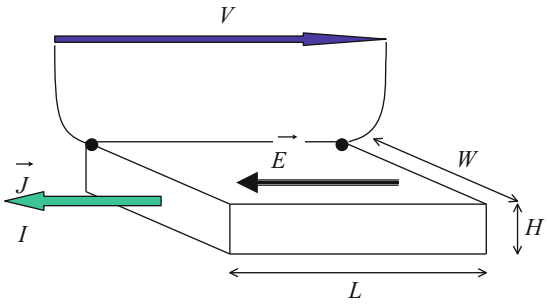


Fig. 8.2 Schematic diagram illustrating the geometry used to illustrate Ohm's law. A voltage is applied across two opposite faces of a rectangular solid and separated by a distance L. This results in an electric field and a current density perpendicular to these two faces

In this configuration, the electrical current I is equal to the magnitude of the current density multiplied by the area WH, i.e., $I = \mathrm{WHJ_{drift}}$. The voltage V is equal to the magnitude of the electric field strength $\left|\vec{E}\right|$ multiplied by the length L of solid considered, i.e., $V = L\left|\vec{E}\right|$. We therefore get successively:

$$\begin{aligned} I &= \mathrm{WHJ^{drift}} = \mathrm{WH}\sigma\left|\vec{E}\right| \\ &= \frac{\mathrm{WH}}{L}\sigma L\left|\vec{E}\right| = \frac{\mathrm{WH}}{L}\sigma V \end{aligned} \tag{8.9}$$

We thus recognize the relation:

$$I = \frac{V}{R} \tag{8.10}$$

where:

$$R = \frac{L}{\mathrm{WH}}\frac{1}{\sigma} \mathrm{or} R = \frac{L}{A}\rho \tag{8.11}$$

and $A = \mathrm{WH}$ is the area of the surface perpendicular to and traversed by the electrical current flow. The quantity R is called the resistance of slab of solid considered. This expression relates a macroscopic quantity (resistance) to an internal property of the solid (resistivity).

8.2.2 The Case of Semiconductors

So far, the discussion has been general and valid for any solid that contains mobile charge carriers. In the case of semiconductors, a few modifications to the previous results need to be made.

A semiconductor has two types of charge carriers which can contribute to the electrical conduction: electrons in the conduction band and holes in the valence band. There are thus two separate contributions to the drift current:

$$\vec{J}^{\text{drift}} = \vec{J}_{\text{e}}^{\text{drift}} + \vec{J}_{\text{h}}^{\text{drift}}$$

where each of the $\vec{J}_{\text{e}}^{\text{drift}}$ and $\vec{J}_{\text{h}}^{\text{drift}}$ is expressed through Eq. (8.7) using the carrier concentrations n and p, mobilities μ_{e} and μ_{h}, and effective masses m_{e} and m_{h} of the electron and the hole, respectively, in the semiconductor considered. Note that, unlike electrons, the holes flow in the same direction as the electric field, because of their positive charge. We thus obtain:

$$\begin{cases} \vec{v}_{\text{e}}^{\text{drift}} = -\mu_{\text{e}} \vec{E} \\ \vec{v}_{\text{h}}^{\text{drift}} = +\mu_{\text{h}} \vec{E} \end{cases} \text{and} \begin{cases} \vec{J}_{\text{e}}^{\text{drift}} = -nq\, \vec{v}_{\text{e}}^{\text{drift}} \\ \vec{J}_{\text{h}}^{\text{drift}} = +pq\, \vec{v}_{\text{h}}^{\text{drift}} \end{cases}. \tag{8.12}$$

The total drift current density can then be written as:

$$\begin{cases} \vec{J}^{\text{drift}} = \sigma \vec{E} \\ \sigma = q(n\mu_{\text{e}} + p\mu_{\text{h}}) \end{cases} \tag{8.13}$$

The typical room temperature conductivity in metals is $(0.1 \sim 3) \times 10^4$ S·cm^{-1}, while the conductivity in semiconductors depends on the carrier concentrations and therefore the doping level, as discussed in Chap. 7.

The conductivity in semiconductors depends much more strongly on the temperature than that in metals. This is because in semiconductors, at a temperature of 0 K, the Fermi energy lies within the forbidden gap (Fig. 5.11) and there is no electron in the conduction band (and thus no hole in the valence band) as the Fermi-Dirac distribution is strictly equal to zero there (Fig. 5.12). Remember that a full band does not carry current. By increasing the temperature, it is therefore possible to increase the concentrations of electrons in the conduction band, holes in the valence band, and enhance electrical conductivity as the Fermi-Dirac distribution is not strictly equal to zero any more. By contrast, in metals, the Fermi energy lies within the conduction band which is thus partially filled (Fig. 5.11), and an increase in temperature will not significantly affect the concentration of electrons in it.

8.3 Carrier Mobility in Solids

The mobility of electrons is controlled by two physical parameters: one is the effective mass and the other is the relaxation time. In Chap. 5, we have seen what determines the effective mass of a charge. Let us now consider the momentum lifetime. The scattering processes which determine the momentum lifetime of solids can be classified into two categories: (a) elastic scattering processes and (b) inelastic scattering processes. In category (a), the carrier changes its momentum but not its energy. Any break in the translational symmetry of the solid will give rise to elastic scattering, and in particular this includes the presence of impurity potentials, defects

interfaces, and dislocations, but there are also the deviations from periodic order caused by lattice vibrations: the electron-phonon interactions. The former contribute to category (a), and the latter involve energy exchange with the lattice and are in category (b). In category (b) the carrier changes both momentum and energy. An inelastic phonon-induced scattering process is allowed if it satisfies both the momentum and energy conservation conditions which are, respectively:

$$\vec{k'} = \vec{k} \pm \vec{q}$$
$$E\left(\vec{k'}\right) = E\left(\vec{k}\right) \pm \hbar\omega\left(\vec{q}\right)$$

where $E\left(\vec{k'}\right)$ is the energy of the particle after the scattering process and $\hbar\omega\left(\vec{q}\right)$is the energy of the phonon absorbed or emitted.

We have seen in Chap. 6 that a solid will in general have two types of phonons, so there are also two types of electron-phonon scattering processes. These are the electron-acoustic and electron-optical phonon scattering processes. The acoustic scattering occurs in all solids, but optic phonon scattering can only take place when there are optic modes in the system. The strength of the electron-acoustic and electron-optical coupling determines the efficiency, or the rate at which a carrier with a given momentum $\vec{k}$ is scattered into a momentum state $\vec{k'}$ via a phonon. In III–V semiconductors with polar modes, the electron optic coupling is an efficient process and is the most important mechanism by which hot carriers relax their excess energy when they have enough energy to emit an optic phonon. An electron can also absorb an optic phonon, but this is only possible if a sufficient number is thermally excited. The rate of optic phonon absorption increases therefore with temperature, following essentially the Bose-Einstein distribution law of phonon occupation. When more than one scattering process is contributing, the sum must be taken. This is done by summing the lifetimes in parallel so that the shortest time dominates. The total lifetime τ is thus given by the sum $\frac{1}{\tau} = \frac{1}{\tau_{\text{el}}} + \frac{1}{\tau_{\text{op}}} + \frac{1}{\tau_{\text{ac}}}$where the terms denote the inverse of the elastic, optic, and acoustic scattering lifetimes, respectively. The temperature dependence of the mobility in different materials is not simple to summarize, and the reader is referred to the specialized textbooks by Ridley and Sze. The physics of the situation however is as follows: at very low temperatures, the phonon modes freeze out and thermal velocities are low, the inelastic lifetimes therefore increase as we go down in temperature, and eventually elastic processes dominate. Elastic scattering processes can however be weakly dependent on temperature and will remain finite even at zero temperature creating a finite resistance unless the material becomes a superconductor at some stage. Elastic scattering can take place from neutral defects, and most effectively also from charged ionized defects and impurities. The state of ionization of an impurity will in general be a function of temperature, as we saw when we discussed doped semiconductors (see Sect. 7.6). This means that elastic scattering processes in doped semiconductors will in general have both strong temperature-dependent and weak temperature-dependent components. Here are a few typical measured bulk values

(see also Appendix A4) of the room temperature ($T = 300$ K) mobilities of some important semiconductors: Si electrons, 1500 cm^2/Vs; Si holes, 450 cm^2/Vs; GaAs electrons, 8500 cm^2/Vs; GaAs holes, 400 cm^2/Vs; InAs electrons, 33,000 cm^2/Vs; and InAs holes, 460 cm^2/Vs. From the example in the text, we calculated the mobility of Cu, which is a good metal, to be ~20 cm^2/Vs. This is typical for good metals and interestingly lower than for many semiconductors.

8.4 Hall Effect

At the end of the nineteenth century, physicists knew that if a metal wire carrying an electrical current was placed in a magnetic field, it experienced a force. The origin of this force was not known. In 1879, E.H. Hall tried to prove that this force was exerted only on the mobile charges (electrons) in the wire. By doing so, he conducted an experiment where an electrical current was run through a fixed conductor perpendicularly to a magnetic field.

Let us consider the Hall effect experiment geometry illustrated in Fig. 8.3. An electrical current, with current density $\vec{J}$ in the x-direction, is run through a parallelepiped-shaped solid. A magnetic induction or flux density $\vec{B}$ is directed perpendicularly to the current, in the z-direction. The movement of holes and electrons is shown in Fig. 8.3 as well.

8.4.1 *P*-Type Semiconductor

Let us now assume that the solid only contains one type of charge carriers and that they are holes. With the electrical current in the *(+x)*-direction, a hole moves also in the x-direction with a velocity $\vec{v}_h$, as shown in Fig. 8.3 . At the same time, it is subject to the Lorentz force equal to:

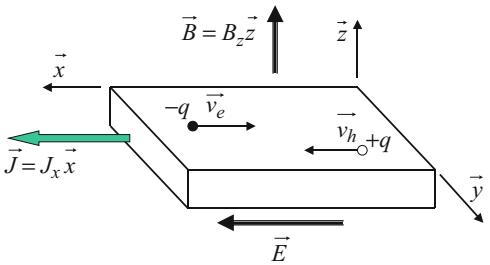


Fig. 8.3 Geometry used for a Hall effect experiment. A uniform electric field strength is applied inside a solid in the x-direction (e.g., by applying a voltage across the solid), which results in an electric current in the same direction. The movement of holes and electrons in the solid is shown. The solid is immersed in a magnetic induction which is directed in the z-direction, perpendicularly to this electric field

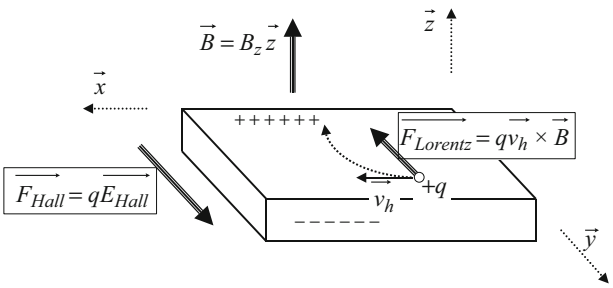


Fig. 8.4 Motion of a hole in the Hall effect experiment. Under the influence of the Lorentz force, the motion of holes is deviated in the *y*-direction toward one side of the solid which then becomes positively charged through the accumulation of holes. The opposite side of the solid therefore becomes negatively charged. This gives rise to an additional electric field which is directed in the y-direction

$$\vec{F}_{\text{Lorentz}} = q\,\vec{v}_{\text{h}} \times \vec{B} \tag{8.14}$$

which is in the *y*-direction. If the sample was without limits, the hole would exhibit a cyclotron (circular) motion around an axis parallel to $\vec{B}$. In the case of a finite size solid as the one shown in Fig. 8.4, holes would accumulate on one of its sides to create a surplus of positive charges. At the same time, negative charges would appear on the opposite side from the deficiency of holes there. This separation of charges results in an electric field strength $\vec{E}_{\text{Hall}}$, called Hall electric field and shown in Fig. 8.4, which drives holes in the *y*-direction and is opposite to the Lorentz force.

At equilibrium, the Lorentz force and the force due to the Hall electric field must balance each other. This can be expressed mathematically as:

$$\vec{0} = \vec{F}_{\text{h, Lorentz}} + \vec{F}_{\text{Hall}} = q\,\vec{v}_{\text{h}} \times \vec{B} + q\,\vec{E}_{\text{Hall}} \tag{8.15}$$

The Hall electric field strength is thus:

$$\vec{E}_{\text{Hall}} = -\,\vec{v}_{\text{h}} \times \vec{B} \tag{8.16}$$

The component of the Hall electric field strength in the *y*-direction (i.e., $\vec{E}_{\text{Hall}} = (E_{\text{Hall}})_y\,\vec{y}$) in the geometry shown in Fig. 8.4 is:

$$(E_{\text{Hall}})_y = (v_{\text{h}})_x B_z > 0 \tag{8.17}$$

From Eq. (8.12), we get:

$$J_x = pq(v_{\text{h}})_x$$

where p is the hole concentration in the solid, and we can rewrite Eq. (8.18) as:

$$(E_{\text{Hall}})_y = \frac{J_x}{pq} B_z \tag{8.18}$$

This expression contains macroscopic quantities which are characteristic of the material (p), parameters of the experiments (J and B), and quantities which are experimentally measured (E_{Hall}). Through this relation, we can easily extract properties characteristic of the materials from experiments. It is a common practice to introduce the Hall constant given by:

$$R_{\text{H}} = \frac{(E_{\text{Hall}})_y}{J_x B_z} = \frac{1}{pq} > 0 \tag{8.19}$$

The Hall constant therefore yields a direct measure of the hole concentration in the solid. We can define a hole Hall mobility as:

$$\mu_{\text{H,h}} = \sigma R_{\text{H}} \tag{8.20}$$

This Hall mobility has the same units as the drift mobility encountered in Eq. (8.12) in Sect. 8.2, i.e., $\text{cm}^2 \cdot \text{V}^{-1} \cdot \text{s}^{-1}$. However, it differs from the drift mobility by a factor, called the Hall factor, which is determined by the temperature and the types of scattering involving the charge carriers. Experimentally, this factor is taken to be equal to unity and only one mobility is considered. This can be illustrated by the fact that one can arrive at Eq. (8.20) from Eq. (8.19) by using the expression in Eq. (8.13) applied to holes only.

8.4.2 *N*-Type Semiconductor

In the case of a solid which contains only electrons as the mobile charge carriers, a similar analysis can be conducted. The motion of an electron in the Hall effect experiment is shown in Fig. 8.5. We can see that the electrons are deflected in the same direction as the holes in Fig. 8.6.

However, because electrons have a negative charge, the Hall electric field is in the opposite direction in comparison to the one from holes:

$$\vec{0} = \vec{F}_{\text{e,Lorentz}} + \vec{F}_{\text{Hall}} = -q\,\vec{v}_{\text{e}} \times \vec{B} - q\,\vec{E}_{\text{Hall}} \tag{8.21}$$

The Hall electric field strength is thus:

$$\vec{E}_{\text{Hall}} = -\vec{v}_{\text{e}} \times \vec{B} \tag{8.22}$$

The component of the Hall electric field strength in the y-direction (i.e., $\vec{E}_{\text{Hall}} = (E_{\text{Hall}})_y\,\vec{y}$) in the geometry shown in Fig. 8.5 is:

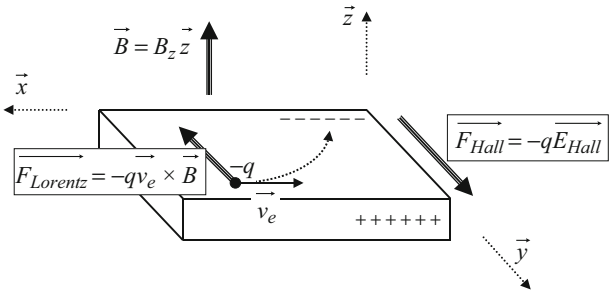


Fig. 8.5 Motion of an electron in the Hall effect experiment. Under the influence of the Lorentz force, the motion of electrons is deviated in the y-direction toward one side of the solid which then becomes negatively charged through the accumulation of electrons. The opposite side of the solid therefore becomes positively charged. This gives rise to an additional electric field which is directed in the y-direction

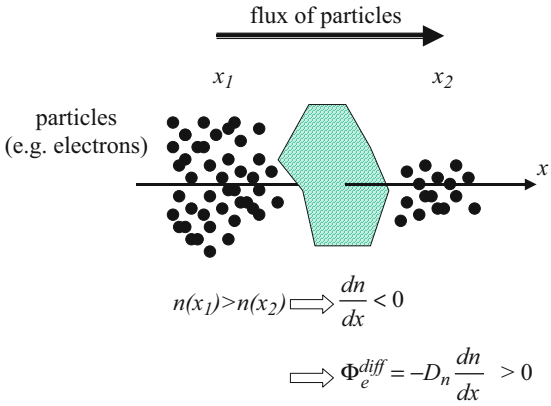


Fig. 8.6 Diffusion of particles (e.g., electrons) in a one-dimensional model. An imaginary surface with a unit area is considered, such that the concentration of particles on one side is larger than on the other side. The diffusion process is characterized by the flux of particles spontaneously passing through the imaginary surface per unit time

$$(E_{\text{Hall}})_y = (v_e)_x B_z < 0 \tag{8.23}$$

because $(v_e)_x < 0$. From Eq. (8.12), we have:

$$J_x = -nq(v_e)_x$$

and we can rewrite Eq. (8.23) as:

$$(E_{\text{Hall}})_y = -\frac{J_x}{nq} B_z \tag{8.24}$$

This expression is similar to Eq. (8.18), and the Hall constant defined in Eq. (8.19) becomes:

$$R_{\mathrm{H}} = \frac{(E_{\mathrm{Hall}})_y}{J_x B_z} = -\frac{1}{nq} < 0 \tag{8.25}$$

Here again, we see that the Hall constant yields the electron concentration in the solid. Moreover, it is negative, whereas it was positive when holes were the only charge carrier. The Hall constant is therefore a good method to determine if a semiconductor is *p*-type or *n*-type. The electron Hall mobility given by Eq. (8.20) is now transformed into:

$$\mu_{\mathrm{H,e}} = \sigma |R_{\mathrm{H}}| > 0 \tag{8.26}$$

Similar to the previous case, the electron Hall mobility is usually taken equal to the electron drift mobility.

8.4.3 Compensated Semiconductor

In a compensated semiconductor, both types of dopants are simultaneously present in the material. Since the electrons and holes released by the doping can recombine, a decrease of the free carriers' concentration can be observed. Adding *p*-type impurities to an *n*-doped system will therefore reduce the electron concentration and vice versa. The charged impurities are still there, having transferred the charge to each other (donor to acceptor) rather than to the bands. It is possible in this way to increase the resistance of doped systems by adding the opposite type of dopant. This can be very useful when ion implantation is used to dope a material, because with ions, one can in principle achieve a high degree of spatial resolution and select the depth of implantation. The ion beam can also be focused to compensate the local doping and thus produce submicron devices.

8.4.4 Hall Effect with Both Types of Charge Carriers

When both electrons and holes are contributing to the transport process, the calculation of the Hall coefficient is somewhat more complicated. Both types of carriers will contribute to the Hall effect in an intrinsic material, for example, or when light is photo-exciting pairs, or when electrons and holes are injected using different types of source drain electrode materials. The derivation of R_{H} is however straightforward and can be done by using the Newton law with the Lorentz force for both carriers:

$$\begin{aligned} m_{\mathrm{e}} \frac{dv_x}{dt} + m_{\mathrm{e}} v_x \frac{1}{\tau} &= -qE_x - qv_y B \\ m_{\mathrm{e}} \frac{dv_y}{dt} + m_{\mathrm{e}} v_y \frac{1}{\tau} &= -qE_y - qv_x B \end{aligned} \tag{8.27}$$

in the presence of electric fields, E_x, E_y, and a magnetic field B. Similar equation can be written down for holes, except that $q \rightarrow -q$. The steady-state velocities are obtained by assuming that the velocity no longer changes with time, i.e., by putting the acceleration term equal to zero. The products μB are very small under typical measurement conditions, so if we ignore even smaller terms of order B^2, we can write Eq. (8.27) as:

$$\begin{aligned} v_y^{\mathrm{e}} &= -\mu_{\mathrm{e}} E_y - \mu_{\mathrm{e}}^2 B E_x \\ v_y^{\mathrm{h}} &= \mu_{\mathrm{h}} E_y - \mu_{\mathrm{h}}^2 B E_x \end{aligned} \tag{8.28}$$

The above equations can be related to the total current J_y giving:

$$J_y = nq\mu_{\mathrm{e}}\left(E_y + \mu_{\mathrm{e}} E_x B\right) + pq\mu_{\mathrm{h}}\left(E_y - \mu_{\mathrm{h}} B E_x\right) \tag{8.29}$$

Under equilibrium condition, i.e., when the current $J_y = 0$, the ratio of the components of the electric field is such that:

$$\frac{E_y}{E_x} = \left\{\frac{p\mu_h^2 - n\mu_e^2}{n\mu_e + p\mu_h}\right\} B \tag{8.30}$$

and the Hall constant is now given by:

$$R_H = \frac{1}{q}\,\frac{p\mu_h^2 - n\mu_e^2}{\left(p\mu_h + n\mu_e\right)^2} \tag{8.31}$$

where p and n are the hole and electron concentrations and μ_{h} and μ_{e} are the hole and electron mobilities, all of which are positive parameters. The Hall mobility is the combination of the mobilities of the electrons and holes and given by:

$$\mu_{\mathrm{H}} = \sigma |R_{\mathrm{H}}| = \left|\frac{p\mu_{\mathrm{h}}^2 - n\mu_{\mathrm{e}}^2}{p\mu_{\mathrm{h}} + n\mu_{\mathrm{e}}}\right| \tag{8.32}$$

8.5 Charge Carrier Diffusion

In an inhomogeneous solid, certain regions may exhibit more electrons or holes than other regions. These will then migrate from the high concentration areas to the low concentration areas. This is a universal and natural phenomenon, called diffusion. This process is due to an imbalance in the thermodynamic chemical potential. One may picture the diffusion process as a drop of ink in a glass of clear water which slowly spreads in the entire volume of water. Because electrons and holes are charge carriers, their diffusion generates an electrical current, which is very important in many semiconductor devices.

8.5.1 Diffusion Currents

In this section, we will describe a simple one-dimensional model for the diffusion of electrons and holes in a semiconductor. Let us assume the electron concentration $n(x)$ is not uniform in the x-direction, as schematically illustrated in Fig. 8.6 .

The diffusion process is mathematically described by Fick's first law of diffusion which says that the flux, i.e., the number of electrons passing per unit time a unit area surface perpendicular to the x-direction, is given by:

$$\Phi_{\mathrm{e}}^{\mathrm{diff}} = -D_{\mathrm{n}}\frac{dn}{dx} \tag{8.33}$$

where D_{n} is called the diffusion coefficient or diffusivity and has the units of $\mathrm{cm}^2 \cdot \mathrm{s}^{-1}$. We use the subscript "n" to identify that this is the diffusivity for electrons. The negative sign in this expression means that the flux of electrons is in the direction opposite to the gradient (or slope) of concentration, as illustrated in Fig. 8.6.

Using a similar approach as for the electrical drift process in Sect. 8.2 to count the number of electrons that pass the unit area surface in Fig. 8.6 per unit time, we can extract the electron diffusion velocity $v_{\mathrm{h}}^{\mathrm{diff}}$:

$$\Phi_{\mathrm{e}}^{\mathrm{diff}} = n v_{\mathrm{e}}^{\mathrm{diff}} \tag{8.34}$$

which leads to the relation:

$$v_{\mathrm{e}}^{\mathrm{diff}} = -D_{\mathrm{n}}\frac{1}{n}\frac{dn}{dx} \tag{8.35}$$

The movement of these electrons creates an electrical current. The diffusion current density of electrons is then determined from Eq. (8.12):

$$J_{\mathrm{e}}^{\mathrm{diff}} = -nqv_{\mathrm{e}}^{\mathrm{diff}} = +qD_{\mathrm{n}}\frac{dn}{dx} \tag{8.36}$$

Similar relations to Eqs. (8.35) and (8.36) can be obtained for the diffusion of holes:

$$v_{\mathrm{h}}^{\mathrm{diff}} = -D_{\mathrm{p}}\frac{1}{p}\frac{dp}{dx} \tag{8.37}$$

$$J_{\mathrm{h}}^{\mathrm{diff}} = -pqv_{\mathrm{h}}^{\mathrm{diff}} = -qD_{\mathrm{p}}\frac{dp}{dx} \tag{8.38}$$

where p is the concentration of holes. Note that there is a sign change from Eqs. (8.36), (8.37), and (8.38) which is due to the positive charge of the hole. There is no such sign change from Eqs. (8.35), (8.36), and (8.37), because the origin of the diffusion process is not dependent on the electrical charge.

8.5.2 Einstein Relations

The drift and the diffusion of electrons and holes are intimately related processes, because both contribute to the observed electrical current in a semiconductor.

Let us continue on our simple one-dimensional model and consider a finite size solid onto which a uniform external electric field of strength $\vec{E} = E\,\vec{x}$ is applied. As a result, the electrons will be drifting to one side of the solid, and a concentration gradient will be achieved. These electrons will then start to diffuse in the direction opposite to this electrical drift until a balance is reached.

The drift current density is given by Eq. (8.12) $J_{\mathrm{e}}^{\mathrm{drift}} = nq\mu_{\mathrm{e}}E$, while the electrical diffusion current density is given by Eq. (8.36) $J_{\mathrm{e}}^{\mathrm{drff}} = +qD_{\mathrm{n}}\dfrac{dn}{dx}$. At the thermal equilibrium of this system, the sum of these two current densities:

$$J_{\mathrm{e}}^{\mathrm{drift}} + J_{\mathrm{e}}^{\mathrm{diff}} = nq\mu_{\mathrm{e}}E + qD_{\mathrm{n}}\frac{dn}{dx} \tag{8.39}$$

must be equal to zero, i.e.:

$$nq\mu_e E + qD_n\frac{dn}{dx} = 0 \tag{8.40}$$

This first-order differential equation can be rewritten as:

$$\frac{dn}{dx} + \frac{\mu_{\mathrm{e}}E}{D_{\mathrm{n}}}n = 0 \tag{8.41}$$

which leads to the solution:

$$n(x) = n(0)\exp\left(-\frac{\mu_{\mathrm{e}}Ex}{D_{\mathrm{n}}}\right) \tag{8.42}$$

where $n(0)$ is the electron concentration at $x = 0$. We see that we obtain an exponential-like distribution for this concentration. However, at thermal equilibrium, this quantity also obeys Boltzmann statistics, which is analogous to the Boltzmann probability distribution we encountered in Chap. 5. For a nondegenerate semiconductor, the electron concentration according to Boltzmann statistics should be given by:

$$n(x) = n(0)\exp\left(-\frac{qEx}{k_{\mathrm{b}}T}\right) \tag{8.43}$$

because qEx is the potential energy of the electron in an electric field strength of magnitude E. Comparing Eqs. (8.42) and (8.43), we obtain the relation:

$$\frac{\mu_e E}{D_n} = \frac{qE}{k_b T}$$

or:

$$\frac{D_n}{\mu_e} = \frac{k_b T}{q} \tag{8.44}$$

A similar relation can be obtained for holes:

$$\frac{D_p}{\mu_h} = \frac{k_b T}{q} \tag{8.45}$$

Equations (8.44) and (8.45) are called the Einstein relations and are valid only for nondegenerate semiconductors. For degenerate semiconductors, we first need to specify the amount of charge in the bands, and a factor involving the Fermi-Dirac integral Eq. (7.13) needs to be included in the above expressions. These relations are important because they provide a mathematical link between the drift and diffusion processes. They are however not always valid. They apply only when there is a small amount of charge in the band edges, which is the most interesting situation in semiconductor technology.

8.5.3 Diffusion Lengths

In the diffusion model considered so far, an electron or a hole can diffuse indefinitely in space. However, in most real case situations, the diffusion range is much more limited.

Let us consider the diffusion of electrons in a one-dimensional semiconductor model, where excess carriers are continuously generated at $x = 0$ and are then allowed to diffuse toward $x \to \infty$. By the term "excess carriers," we mean that an amount of electrons in addition to the thermal equilibrium concentration n_0 is injected into the semiconductor. The mechanisms by which this is achieved will be discussed later in the text. We will denote:

$$\Delta n(x) = n(x) - n_0 \tag{8.46}$$

the excess electron concentration which is a function of position. A possible shape for $\Delta n(x)$ is shown in Fig. 8.7.

During the diffusion process, an electron will experience recombination, i.e., they will not travel in space indefinitely but will be stopped, for example, when it encounters a hole (remember that a hole is an allowed state vacated by an electron) or when it gets trapped by a defect in the semiconductor crystal (e.g., an ionized donor which is positively charged).

The recombination mechanisms are numerous and diverse. However, it is possible to mathematically express their effects in a simple manner. For this, we introduce

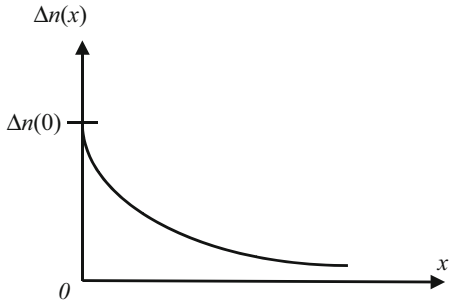


Fig. 8.7 Excess electron concentration in a one-dimensional model. The excess concentration decreases, as it gets deeper into the material as a result of recombination. The decrease has an exponential dependence

a characteristic time, τ_n, called the electron recombination lifetime, such that the recombination rate of an electron at a location where there is an excess $\Delta n(x)$ of electrons is given by:

$$R(x) = \frac{\Delta n(x)}{\tau_n} \tag{8.47}$$

This quantity has the units of $\mathrm{cm}^{-3}\cdot\mathrm{s}^{-1}$ and expresses the change in the excess carrier concentration per unit time.

Let us now consider an infinitesimal region of space, located between x_0 and $x_0 + dx$, as illustrated in Fig. 8.8. This region experiences an influx and an outflux of electrons, denoted, respectively, $\left(\Phi_e^{diff}\right)_{in}$ and $\left(\Phi_e^{diff}\right)_{out}$ and shown in Fig. 8.8.

If $\left(\Phi_e^{diff}\right)_{in} > \left(\Phi_e^{diff}\right)_{out}$, there is a net influx or accumulation of electrons, but if $\left(\Phi_e^{diff}\right)_{out} > \left(\Phi_e^{diff}\right)_{in}$, there is a net outflux or depletion of electrons in this region. Under steady-state conditions, there must not be a never-ending accumulation or depletion of electrons. The influx of electrons must therefore be equal to the sum of the outflux of electrons and the number of electrons recombining within this region. The later quantity is equal to $R(x_0)$ multiplied by the width of the region dx, because we can assume that the function $R(x)$ does not vary too much over a narrow width dx around the point x_0. Numerically, this translates into:

$$\left(\Phi_e^{diff}\right)_{in} = \left(\Phi_e^{diff}\right)_{out} + R(x_0)dx \tag{8.48}$$

From Eq. (8.33), we can write:

$$\left(\Phi_e^{diff}\right)_{in} = -D_n\left(\frac{dn}{dx}\right)_{x=x_0} \text{ and } \left(\Phi_e^{diff}\right)_{out} = -D_n\left(\frac{dn}{dx}\right)_{x=x_0+dx}$$

But, from Eq. (8.46), we easily see that $\frac{dn}{dx} = \frac{d(\Delta n)}{dx}$ and therefore:

$$\begin{cases} \left(\Phi_e^{diff}\right)_{in} = -D_n\left(\frac{d(\Delta n)}{dx}\right)_{x=x_0} \\ \left(\Phi_e^{diff}\right)_{out} = -D_n\left(\frac{d(\Delta n)}{dx}\right)_{x=x_0+dx} \end{cases} \tag{8.49}$$

Equation (8.48) becomes then:

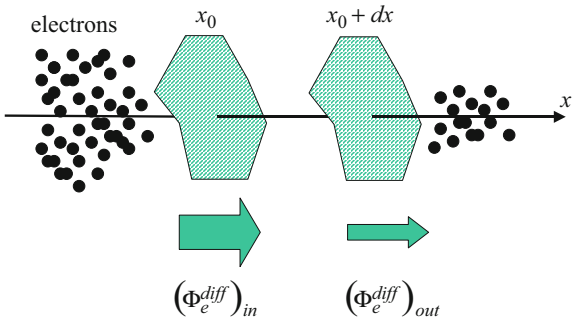


Fig. 8.8 Schematic of the influx and outflux of electrons in a region of space, in a one-dimensional model. In this experiment, the region between the two surfaces located at x_0 and $x_0 + dx$ is considered. This experiment is aimed at determining the net change in carrier concentration in it as a result of the diffusion of particles and their recombination

$$-D_{\mathrm{n}}\left(\frac{d(\Delta_{\mathrm{n}})}{dx}\right)_{x=x_0} = -D_{\mathrm{n}}\left(\frac{d(\Delta n)}{dx}\right)_{x=x_0+dx} + R(x_0)dx$$

which can be rewritten as:

$$D_{\mathrm{n}}\frac{\left(\frac{d(\Delta n)}{dx}\right)_{x=x_0+dx} - \left(\frac{d(\Delta n)}{dx}\right)_{x=x_0}}{dx} = R(x_0)$$

At the limit of $dx \to 0$, i.e., an infinitesimal region, the left-hand side expression becomes the derivative of $\frac{d(\Delta n)}{dx}$ evaluated at $x = x_0$, i.e.:

$$D_{\mathrm{n}}\left(\frac{d^2(\Delta n)}{dx^2}\right)_{x=x_0} = R(x_0) \tag{8.50}$$

This relation is valid for any arbitrarily chosen position x_0, which means that the following equation must be satisfied:

$$D_{\mathrm{n}}\frac{d^2(\Delta n)}{dx^2} = R(x)$$

Equating to Eq. (8.47), we get the differential equation that governs the shape of the excess electron concentration $\Delta n(x)$:

$$D_{\mathrm{n}}\frac{d^2(\Delta n)}{dx^2} = \frac{\Delta n}{\tau_{\mathrm{n}}} \tag{8.51}$$

This equation can be rewritten as:

$$\frac{d^2(\Delta n)}{dx^2} - \frac{\Delta n}{D_{\mathrm{n}}\tau_{\mathrm{n}}} = 0 \tag{8.52}$$

From this expression, we can easily see that the quantity $D_n\tau_n$ has the same dimension as the square of a distance. We can then define a distance L_n, called diffusion length, for electrons, given by:

$$L_n = \sqrt{D_n\tau_n} \tag{8.53}$$

The solution to Eq. (8.52) then has the general form:

$$\Delta n(x) = Ae^{\frac{x}{L_n}} + Be^{\frac{-x}{L_n}} \tag{8.54}$$

Here A and B are constants and are determined from boundary conditions. For example, let us assume the sample is delimited by $x = 0$ and $x \to \infty$, and that is thick enough so that all the excess electrons have been recombined before they reach its limit: $\Delta n \to 0$ when $x \to \infty$ as shown in Fig. 8.7. We thus have:

$$\Delta n(x) = \Delta n(0)e^{-\frac{x}{L_n}} \tag{8.55}$$

From this expression, we see the significance of the diffusion length in determining the spatial distribution of the electrons in the diffusion process as the characteristic length of path that a particle travels before recombining.

A similar diffusion length can be determined for holes and is given by:

$$L_p = \sqrt{D_p\tau_p} \tag{8.56}$$

where τ_p is the hole recombination lifetime.

Example

Q Assuming that in n-type silicon the characteristic time for the minority carriers (holes) is $\tau_p = 2 \times 10^{-10}$ s, estimate the diffusion length of these minority carriers at 300 K.

A The diffusion length is given by $L_p = \sqrt{D_p\tau_p}$. From the Einstein relations, we can determine the diffusion coefficient $D_p = \frac{k_b T\mu_h}{q}$. With the hole mobility in silicon being about $\mu_h = 450$ cm^2/Vs, we get:

$$\begin{aligned} L_p &= \sqrt{\frac{k_b T\mu_h}{q}\tau_p} \\ &= \sqrt{\frac{\left(1.38066 \times 10^{-23}\right) \times 300 \times \left(450 \times 10^{-4}\right)}{1.60218 \times 10^{-19}} \times 2 \times 10^{-10}} \\ &= 0.48\,\mu m \end{aligned}$$

8.6 Carrier Generation and Recombination Mechanisms

In the previous section, we briefly talked about excess carriers and their recombination. We also introduced a single recombination lifetime τ in order to avoid a detailed description of all the recombination processes.

Excess of carriers can exist when the semiconductor is not in its equilibrium state, as a result of additional energy that it received from phonons (heat), photons (light), or an electric field, for example. In a recombination process, the amount of excess carriers is reduced, and the excess energy is transferred or released.

In this section, we will discuss the four most important recombination mechanisms encountered in semiconductors, including direct band-to-band, Shockley-Read-Hall, Auger, and surface recombination. We will also attempt to express the recombination lifetime in each case in terms of known semiconductor parameters.

We will denote by:

$$\begin{cases} \Delta n(t) = n(t) - n_0 \\ \Delta p(t) = p(t) - p_0 \end{cases} \tag{8.57}$$

the excess electron and hole concentrations, respectively, where n_0 and p_0 are the equilibrium electron and hole concentrations.

It is important, at this time, to clearly distinguish equilibrium state from steady state. A system is said to be under equilibrium if it is not subject to external fields or forces. A system under the influence of external fields or forces is under steady state if the parameters that describe it (e.g., carrier concentrations) do not vary with time.

8.6.1 Carrier Generation

Before discussing the various recombination mechanisms, we must first review how carriers are generated in the first place. There are essentially two major types of generation.

The first one corresponds to the thermal generation of carriers and exists under all conditions, whether in equilibrium or non-equilibrium. The thermal generation rate will be denoted $G_t(T)$ and is expressed in units of $\text{cm}^{-3}\cdot\text{s}^{-1}$.

The other type is the generation resulting from external factors, such as optical absorption, electrical injection, etc. This process occurs only in non-equilibrium situations, and the associated generation rate, denoted G, is called the excess generation rate.

For each generation mechanism, there exists a recombination mechanism which is its counterpart. The generation and recombination of carriers are inverse processes to each other.

8.6.2 Direct Band-to-Band Recombination

In this type of recombination, an electron from the conduction band recombines with a hole in the valence band. This process is best pictured in the *E-k* diagram shown in Fig. 8.9 .

This recombination can be equivalently viewed as an electron which goes from a state in the conduction band to an allowed state in the valence band. This seems natural if we remember that a hole in the valence band is in fact an allowed electronic state that has been *vacated* by an electron. The energy that the electron thus loses is most often released in the form of a photon or light as shown in Fig. 8.9. We say that this is a radiative recombination.

This process is most likely to occur between the minimum of the conduction band and the maximum of the valence band and at the center of the first Brillouin zone where the momenta of the recombining electron and hole are both zero. Direct band-to-band radiative recombination is therefore most likely to occur in direct bandgap semiconductors, such as GaAs.

Let us look at this recombination mechanism in more detail. In the present case, the recombination rate, first introduced in Eq. (8.47), is proportional to both the concentration of electrons in the conduction band n and that of holes in the valence band p because these are the particles that are recombining. We can then write:

$$R = r(T)n(t)p(t) \tag{8.58}$$

where $r(T)$ is the recombination coefficient, which is expressed in units of $\text{cm}^3 \cdot \text{s}$, and T is the temperature.

In a non-equilibrium situation when the excess generation rate is nonzero, the net change in the electron and hole densities is given by:

$$-\frac{dn}{dt} = -\frac{d(\Delta n)}{dt} = R - G - G_t \tag{8.59}$$

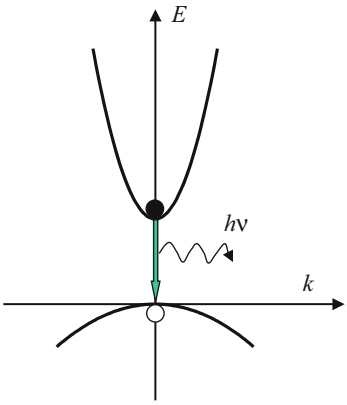


Fig. 8.9 Schematic E-k diagram of a direct band-to-band recombination process. The recombining electron and hole have the same wavevector

where we used the fact that the equilibrium concentration n_0 does not vary with time. At equilibrium, the excess generation rate G is equal to zero; thus the recombination rate must balance the thermal generation rate: $R = G_t$. Since at equilibrium we have $n = n_0$ and $p = p_0$, we can write from Eq. (8.58):

$$G_t = r(T)n_0p_0 \quad \text{or simply} \quad G_t = r(T)n_i^2 \tag{8.60}$$

where n_i is the intrinsic carrier concentration given in Eq. (7.31). From now, we will also omit the temperature dependence of $r(T)$ to simplify the equations.

Let us now consider the relaxation process, which occurs after the external source of generation is removed ($G = 0$). Taking into account Eqs. (8.58) and (8.60), Eq. (8.59) becomes:

$$-\frac{d(\Delta n)}{dt} = r\left[np - n_i^2\right] \tag{8.61}$$

Using Eq. (8.57), we can expand this expression into:

$$-\frac{d(\Delta n)}{dt} = r\left[(n_0 + \Delta n)(p_0 + \Delta p) - n_i^2\right]$$

i.e.:

$$-\frac{d(\Delta n)}{dt} = r\left[n_0p_0 + n_0\Delta p + p_0\Delta n + \Delta n\Delta p - n_i^2\right] \tag{8.62}$$

One obvious simplification can be immediately made in the previous expression as $n_0p_0 = n_i^2$ from Eq. (7.31). For further simplicity, we can assume $\Delta n = \Delta p$, i.e., the concentration of excess electrons is equal to the concentration of excess holes, which seems natural in order to ensure charge neutrality locally in the semiconductor at all times. Equation (8.62) then becomes:

$$-\frac{d(\Delta p)}{dt} = -\frac{d(\Delta n)}{dt} = r[(n_0 + p_0) + \Delta n]\Delta n \tag{8.63}$$

We can successively transform Eq. (8.63) into:

$$-\frac{{}^{d(\Delta n)}/_{dt}}{[(n_0 + p_0) + \Delta n]\Delta n} = r$$

$$\frac{1}{(n_0 + p_0)}\left(\frac{{}^{d(\Delta n)}/_{dt}}{(n_0 + p_0) + \Delta n} - \frac{{}^{d(\Delta n)}/_{dt}}{\Delta n}\right) = r$$

Each of the terms in the left-hand side is a logarithmic derivative. By integrating with respect to time from 0 to t, we get successively:

$$\frac{1}{(n_0+p_0)}\left[\ln\left((n_0+p_0)+\Delta n\right)-\ln\left(\Delta n\right)\right]_0^t = rt$$

$$\frac{1}{(n_0+p_0)}\left[\ln\left(\frac{(n_0+p_0)+\Delta n}{\Delta n}\right)\right]_0^t = rt$$

$$\ln\left(\frac{(n_0+p_0)+\Delta n(t)}{\Delta n(t)}\right)-\ln\left(\frac{(n_0+p_0)+\Delta n(0)}{\Delta n(0)}\right)=r(n_0+p_0)t$$

Taking the exponential on both sides of this last equation, we obtain:

$$\frac{(n_0+p_0)+\Delta n(t)}{\Delta n(t)}=\frac{(n_0+p_0)+\Delta n(0)}{\Delta n(0)}\exp[r(n_0+p_0)t]$$

and solving for $\Delta n(t)$, we get:

$$\Delta p(t)=\Delta n(t)=\frac{(n_0+p_0)\Delta n(0)}{[(n_0+p_0)+\Delta n(0)]\exp[r(n_0+p_0)t]-\Delta n(0)} \tag{8.64}$$

This shows the general form for the change in the excess electron concentration as a function of time. The only parameters of the variation are the equilibrium concentrations n_0 and p_0, the initial excess electron concentration $\Delta n(0)$, and the recombination coefficient $r(T)$. This complicated expression can be drastically simplified in some cases.

For weak excitation levels, i.e., $\Delta n(0) << (n_0 + p_0)$, Eq. (8.64) becomes:

$$\begin{aligned}\Delta n(t) &\approx \frac{(n_0+p_0)\Delta n(0)}{(n_0+p_0)\exp[r(n_0+p_0)t]-\Delta n(0)}\\ &\approx \frac{(n_0+p_0)\Delta n(0)}{(n_0+p_0)\exp[r(n_0+p_0)t]}\end{aligned}$$

or simply:

$$\Delta n(t)\approx \Delta n(0)\exp[-r(n_0+p_0)t] \tag{8.65}$$

and similarly for $\Delta p(t)$:

$$\Delta p(t)\approx \Delta p(0)\exp[-r(n_0+p_0)t] \tag{8.66}$$

By defining a direct band-to-band recombination lifetime for electrons and holes as:

$$\tau_p=\tau_n=\frac{1}{r(n_0+p_0)} \tag{8.67}$$

we obtain:

$$\begin{cases} \Delta n(t) \approx \Delta n(0) e^{-\frac{t}{\tau_n}} \\ \Delta p(t) \approx \Delta p(0) e^{-\frac{t}{\tau_0}} \end{cases} \tag{8.68}$$

This is the same lifetime introduced in Eq. (8.47). Indeed, in the current conditions, we have by using Eqs. (8.59) and (8.68):

$$R - G_t = -\frac{d(\Delta n)}{dt} = \frac{1}{\tau_n} \Delta n(0) e^{-\frac{t}{\tau_n}}$$

or:

$$R - G_t = \frac{\Delta n(t)}{\tau_n} \tag{8.69}$$

which is analogous to Eq. (8.47).

8.6.3 Shockley-Read-Hall Recombination

The previous band-to-band recombination most often occurs in pure semiconductor. When defects or impurities are present in the crystal, which is nearly always the case to some extent, energy levels appear in the bandgap and may participate in the recombination mechanisms. These are called Shockley-Read-Hall (SRH) recombinations, and the energy is not released in the form of a photon but is rather given to the crystal lattice in the form of phonons. Such processes are also sometimes called band-to-impurity recombinations. This is therefore normally a non-radiative recombination step.

In the present model, we consider the steady-state generation and recombination of electrons and holes involving an impurity level, also called recombination center, with an energy E_T in the bandgap, as shown in Fig. 8.10 . Let us assume that electrons and holes are generated at a rate equal to G, which is the excess generation rate of Subsect. 8.6.1.

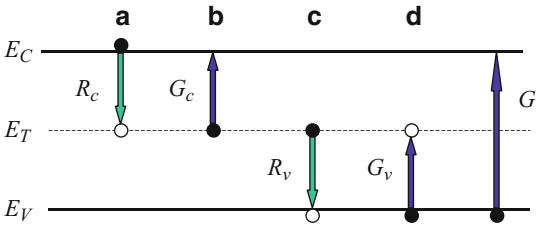


Fig. 8.10 The four possible transitions for an electron involving a recombination center in the bandgap: (**a**) capture of an electron from the conduction band by the center, (**b**) emission of an electron from the center into the conduction band, (**c**) emission of an electron from the center into a vacant state in the valence band, and (**d**) capture of an electron from the valence band by the center

There are four possible electron transitions which can involve this level: (a) the capture of an electron from the conduction band by the center, (b) the emission of an electron from the center into the conduction band, (c) the emission of an electron from the center into a vacant state in the valence band, and (d) the capture of an electron from the valence band by the center. The transition (c) can be equivalently viewed as the capture of a hole by the center and (d) as the emission of a hole from the center into the valence band. Each of these transitions is illustrated in Fig. 8.10.

The recombination of electrons or holes is enhanced by the presence of the impurity level if the probability of transitions (a) and (c) is higher than that of (b) and (d).

If the probability of (a) and (b) is higher than (c) and (d), the impurity level plays more the role of an electron recombination center. If the probability of (c) and (d) is higher than (a) and (b), the impurity level plays more the role of a hole recombination center.

Before analyzing each transition in more detail, let us first assume there is a density N_T of impurity-related states at an energy E_T. At thermal equilibrium, the density of the recombination center states which are occupied by electrons is then given by:

$$N_T f_e(E_T) = \frac{N_T}{\exp\left(\frac{E_T - E_F}{k_b T}\right) + 1} \tag{8.70}$$

where f_e is the Fermi-Dirac distribution given by Eq. (5.28). The density of the recombination center states which are empty of electrons at equilibrium is given by:

$$N_T[1 - f_e(E_T)] = \frac{N_T}{1 + \exp\left(-\frac{E_T - E_F}{k_b T}\right)} \tag{8.71}$$

However, when carriers are transiting through the recombination centers in Fig. 8.10, the density of occupied and empty center states is different from their equilibrium values. We thus introduce a non-equilibrium distribution function f such that the densities of occupied and empty center states are $N_T f$ and $N_T(1 - f)$, respectively. Knowledge of the exact value of this function is not important in analyzing each of the transitions illustrated in Fig. 8.10.

Transition Rates

Let us first discuss the transition (a), i.e., the capture of an electron from the conduction band by the center. The capture rate, or concentration of electrons captured by the center per unit time, is denoted R_c and is expressed in units of cm $^{-3}\cdot s^{-1}$. It must be proportional to the density of electrons in the conduction band n and the density of empty centers $N_T(1 - f)$.

In addition, R_c should also depend on a parameter which describes "how often an electron encounters the recombination center." This parameter is the product $\nu_{th}\sigma_n$ of two quantities: the electron thermal velocity ν_{th} (in units of cm$\cdot s^{-1}$) and the capture cross section σ_n of electrons for this particular recombination center (in units of cm^2).

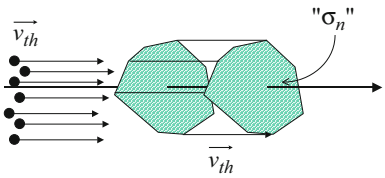


Fig. 8.11 Schematic illustration of the concepts of electron thermal velocity and capture cross section. Using ballistic terminology, the electrons moving with the thermal velocity which would collide with an object having a cross section equal to σ_n are located in the volume delimited by the two shaded surfaces in this figure

These two parameters can be better understood by considering the illustration in Fig. 8.11. It shows that the electrons which have a velocity ν_{th} and which will reach a surface of area σ_n are located in a volume equal to the product $\nu_{th}\sigma_n$ during a unit time.

The electron thermal velocity in a nondegenerate semiconductor is given by:

$$\nu_{th} = \sqrt{\frac{3k_b T}{m}} \tag{8.72}$$

where m is the mass of the electron. The thermal velocity is on the order of 10^7 cm·s^{-1} at room temperature.

The capture cross section of electrons for a recombination center characterizes the interaction between an electron and this center. It corresponds to the effective area around the center that an electron experiences when it is approaching the center. The cross section depends on the type of interaction involved between the electron and the center: the stronger the interaction is, the larger the influence of the capture cross section is. σ_n is usually determined empirically and is on the order of 10^{-15} cm^2. The capture rate R_c in the transition (a) is therefore equal to:

$$R_c = \nu_{th}\sigma_n n N_T (1 - f) \tag{8.73}$$

The emission of an electron from the center into the conduction band, corresponding to transition (b) in Fig. 8.10, is characterized by an emission rate G_c which has the same units as R_c. This quantity is equal to the density of occupied center states $N_T f$ multiplied by the electron emission probability e_n which is a parameter characteristic of the recombination center in the semiconductor:

$$G_c = e_n N_T f \tag{8.74}$$

Because the transitions (c) and (d) are analogous to (a) and (b) but involve holes instead of electrons, we can easily determine the hole capture rate R_v and the hole emission rate G_v from those for electrons Eqs. (8.73) and (8.74).

Indeed, R_v must be proportional to the density of holes in the valence band p, the density of centers which are occupied (by electrons) $N_T f$, the thermal velocity of

holes which is the same as that of electrons given in Eq. (8.72), and the capture cross section of holes σ_p for the center considered:

$$R_v = \nu_{th}\sigma_p p N_T f \tag{8.75}$$

G_v must be equal to the density of center states which are empty (of electrons) $N_T(1-f)$ multiplied by the hole emission probability e_p:

$$G_\nu = e_p N_T(1-f) \tag{8.76}$$

All these expressions for the recombination and emission rates are not independent but must satisfy a number of equations arising from the conservation of electrons and holes. The total number of electrons (or holes) recombined must be equal to the number of electrons (or holes) generated; thus, we can write:

$$\begin{cases} R_c = G_c + G \\ R_v = G_v + G \end{cases} \tag{8.77}$$

Emission Probabilities e_n and e_p

At equilibrium, the excess generation rate G is equal to zero. Moreover, the electron and hole densities are equal n_0 and p_0, respectively, and the distribution function f is equal to $f_e = f_e(E_T)$. All the other parameters remain unchanged. Therefore, by expressing Eq. (8.77) at equilibrium using Eqs. (8.73), (8.74), (8.75), and (8.76), we get:

$$\begin{cases} \nu_{th}\sigma_n n_0 N_T(1-f_e) = e_n N_T f_e \\ \nu_{th}\sigma_p p_0 N_T f_e = e_p N_T(1-f_e) \end{cases}$$

which allow us to extract the electron and hole emission probabilities:

$$\begin{cases} e_n = \nu_{th}\sigma_n n_0 \dfrac{1-f_e}{f_e} \\ e_p = \nu_{th}\sigma_p p_0 \dfrac{f_e}{1-f_e} \end{cases} \tag{8.78}$$

This last set of equations can be simplified by using the expression for the Fermi-Dirac distribution in Eq. (5.28) to obtain:

$$\frac{1-f_e}{f_e} = \exp\left(\frac{E_T - E_F}{k_b T}\right) \tag{8.79}$$

and by using the expressions of n_0 and p_0 given in Eqs. (7.21) and (7.29) for a nondegenerate semiconductor:

$$n_0 \frac{1-f_e}{f_e} = N_c \exp\left(\frac{E_F - E_C}{k_b T}\right) \exp\left(\frac{E_T - E_F}{k_b T}\right)$$
$$= N_c \exp\left(\frac{E_T - E_C}{k_b T}\right)$$

This last quantity can be denoted n_T and would correspond to the electron density in the conduction band if the Fermi energy was equal to the recombination center energy level ($E_F = E_T$):

$$n_T = n_0 \frac{1-f_e}{f_e} = N_c \exp\left(\frac{E_T - E_C}{k_b T}\right) \tag{8.80}$$

A similar expression can be derived for:

$$p_T = p_0 \frac{f_e}{1-f_e} = N_v \exp\left(\frac{E_V - E_T}{k_b T}\right) \tag{8.81}$$

Therefore, Eq. (8.78) is simplified into:

$$\begin{cases} e_n = \nu_{th} \sigma_n n_T \\ e_p = \nu_{th} \sigma_p p_T \end{cases} \tag{8.82}$$

The Non-equilibrium Distribution Function f

The non-equilibrium distribution function, included in the expressions of the transition rates in Eqs. (8.73), (8.74), (8.75), and (8.76), can be determined by eliminating the excess generation rate G in Eq. (8.77). For this, we first calculate the difference between the two expressions in Eq. (8.77):

$$R_c - R_v = G_c - G_v$$

which becomes:

$$\nu_{th} \sigma_n n N_T (1-f) - \nu_{th} \sigma_p p N_T f = e_n N_T f - e_p N_T (1-f)$$

Using Eq. (8.82), we obtain:

$$\nu_{th} \sigma_n n N_T (1-f) - \nu_{th} \sigma_p p N_T f = \nu_{th} \sigma_n n_T N_T f - \nu_{th} \sigma_p p_T N_T (1-f)$$

and, after simplifying by ν_{th} and N_T:

$$\sigma_n n + \sigma_p p_T = f \left\lfloor \sigma_n n + \sigma_p p + \sigma_n n_T + \sigma_p p_T \right\rfloor$$

Thus finally we have:

$$f = \frac{\sigma_n n + \sigma_p p_T}{\sigma_n (n + n_T) + \sigma_p (p + p_T)} \tag{8.83}$$

Recombination Lifetimes

The net recombination rate of electrons from the conduction band is given by the difference between the recombination rate R_c and the generation rate G_c, i.e.:

$$-\frac{d(\Delta n)}{dt} = R_c - G_c \tag{8.84}$$

This quantity is also equal to the net recombination rate of holes from the valence band in view of Eq. (8.77):

$$-\frac{d(\Delta p)}{dt} = R_v - G_v \tag{8.85}$$

Using the non-equilibrium distribution function (Eq. 8.83) and the expressions for R_c, G_c, and e_n in Eqs. (8.73), (8.74) and (8.82), we can calculate successively:

$$\begin{aligned} R_c - G_c &= v_{th}\sigma_n n N_T(1-f) - e_N N_T f \\ &= v_{th}\sigma_n N_T[n-(n+n_T)f] \\ &= v_{th}\sigma_n N_T\left[n-(n+n_T)\frac{\sigma_n n+\sigma_p p_T}{\sigma_n(n+n_T)+\sigma_p(p+p_T)}\right] \\ &= \frac{v_{th}\sigma_n N_T}{\sigma_n(n+n_T)+\sigma_p(p+p_T)}\left[n\sigma_p(p+p_T)-(n+n_T)\sigma_p p_T\right] \\ &= \frac{v_{th}\sigma_n N_T}{\sigma_n(n+n_T)+\sigma_p(p+p_T)}\sigma_p[np-n_T p_T] \end{aligned}$$

From the definitions of n_T and p_T in Eqs. (8.80) and (8.81), we have $n_T p_T = n_i^2$ where n_i is the intrinsic carrier concentration. The previous equation can then be simplified into:

$$R_c - G_c = v_{th}\sigma_n\sigma_p N_T \frac{(np-n_i^2)}{\sigma_n(n+n_T)+\sigma_p(p+p_T)} \tag{8.86}$$

Introducing the excess carriers Δn and Δp as in Eq. (8.57), and still assuming $\Delta n = \Delta p$, we get:

$$R_c - G_c = v_{th}\sigma_n\sigma_p N_T \frac{(n_0+p_0+\Delta n)\Delta n}{\sigma_n(n_0+n_T+\Delta n)+\sigma_p(p_0+p_T+\Delta n)} \tag{8.87}$$

Here we have also used the relation $n_0 p_0 = n_i^2$. This expression can be further simplified by first considering two particular cases.

(i) For low excess carrier concentrations, i.e., weak excitation levels where $\Delta n << n_0$, p_0, and for an n-type semiconductor, where we can assume that n_0 is much higher than p_0, n_T, and p_T, Eq. (8.87) becomes:

$$R_c - G_c \approx v_{th}\sigma_n\sigma_p N_T \frac{(n_0)\Delta n}{\sigma_n(n_0)}$$

which can be rewritten, by taking into account Eq. (8.84):

$$-\frac{d(\Delta n)}{dt} = R_c - G_c \approx v_{th}\sigma_p N_T \Delta n \tag{8.88}$$

From this last expression, we can introduce a recombination lifetime τ_{p_0} such that:

$$-\frac{d(\Delta n)}{dt} \approx \frac{\Delta n}{\tau_{p_0}}$$

i.e.:

$$\tau_{p_0} = \frac{1}{v_{th}\sigma_p N_T} \tag{8.89}$$

Note that the subscript "p" has been used for this lifetime, because it depends on the capture cross section of holes. This corresponds to a lifetime of holes. Therefore, in an *n*-type semiconductor, the excess carrier lifetime approaches that of holes.

(ii) In the second case, still $\Delta n << n_0, p_0$; but for a *p*-type semiconductor this time, where we can assume that p_0 is much higher than n_0, n_T, and p_T, Eq. (8.87) becomes:

$$R_c - G_c \approx v_{th}\sigma_n N_T \Delta n$$

Here again, we can rewrite this as:

$$-\frac{d(\Delta n)}{dt} \approx \frac{\Delta n}{\tau_{n_0}}$$

with:

$$\tau_{n_0} = \frac{1}{v_{th}\sigma_n N_T} \tag{8.90}$$

Here, the suffix "*n*" has been used, because the lifetime depends on the capture cross section of electrons. This corresponds to a lifetime of electrons. Therefore, in a *p*-type semiconductor, the excess carrier lifetime approaches that of electrons. Using the expressions in Eqs. (8.89) and (8.90), we can simplify Eq. (8.87):

$$R_c - G_c = \frac{(n_0 + p_0 + \Delta n)\Delta n}{\tau_{p_0}(n_0 + n_T + \Delta n) + \tau_{n_0}(p_0 + p_T + \Delta n)} \tag{8.91}$$

From Eqs. (8.84) and (8.85), we can write:

$$-\frac{d(\Delta n)}{dt} = -\frac{d(\Delta p)}{dt} = \frac{(n_0 + p_0 + \Delta n)\Delta n}{\tau_{p_0}(n_0 + n_T + \Delta n) + \tau_{n_0}(p_0 + p_T + \Delta n)} \tag{8.92}$$

We can now introduce the Shockley-Read-Hall recombination lifetime $\tau_n = \tau_p$ such that:

$$-\frac{d(\Delta n)}{dt} = -\frac{d(\Delta p)}{dt} = \frac{\Delta p}{\tau_p} = \frac{\Delta n}{\tau_n}$$

i.e.:

$$\tau_n(t) = \tau_p(t) = \frac{\tau_{p_0}(n_0 + n_T + \Delta n) + \tau_{n_0}(p_0 + p_T + \Delta n)}{(n_0 + p_0 + \Delta n)} \tag{8.93}$$

which becomes independent of time for weak excitation levels $\Delta n << n_0, p_0$:

$$\tau_n = \tau_p = \frac{\tau_{p_0}(n_0 + n_T) + \tau_{n_0}(p_0 + p_T)}{(n_0 + p_0)} \tag{8.94}$$

From this relation, we can easily find the two previous particular cases, i.e., for an n-type semiconductor, $\tau_n = \tau_p = \tau_{p_0}$, and for a p-type semiconductor, $\tau_n = \tau_p = \tau_{n_0}$.

8.6.4 Auger Band-to-Band Recombination

Unlike the direct band-to-band or the SRH processes, in the Auger band-to-band, or simply Auger recombination, the energy that is released when an electron recombines with a hole is transferred to a third particle, an electron in the conduction band or a hole in the valence band. This carrier particle is called an Auger electron or Auger hole. The energy that this third particle acquires is subsequently released in the form of heat or phonons into the lattice. Auger recombination is an intrinsic non-radiative mechanism which is more effective at higher temperatures and for smaller bandgap semiconductors. This recombination mechanism occurs most often in doped direct bandgap semiconductors.

There are three possible Auger recombination mechanisms, depending on what type of Auger carrier is excited and where it is excited. These are illustrated in Fig. 8.12.

The first process, shown in Fig. 8.12a, is called a CHCC process to indicate that an electron from the conduction band (C) recombines with a hole in the valence band (H) to lead to the excitation of another electron which remains in the conduction

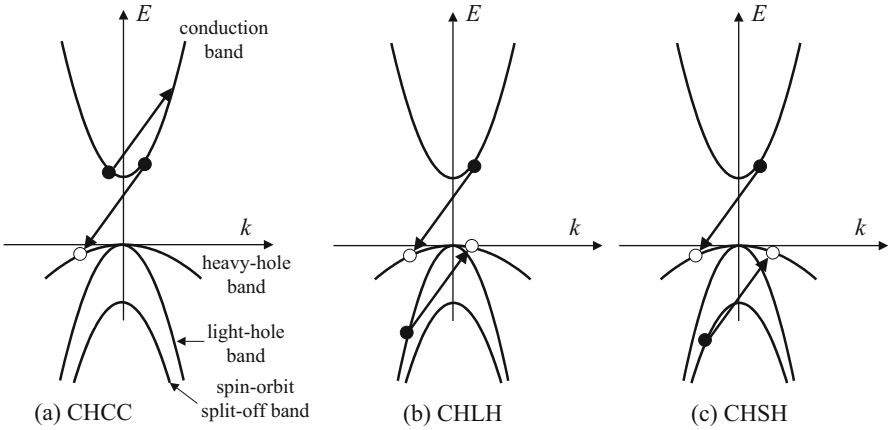


Fig. 8.12 Auger recombination process semiconductors. The energy released through the recombination of an electron in the conduction band and a hole in the valence band is yielded to: (**a**) another electron in the conduction band which is then excited to a higher state in the band, (**b**) an electron in the LH band which is excited to a vacant electronic state in the HH band, (**c**) an electron in the split-off band which is excited to a vacant electronic state in the HH band

band (CC). In the case of an Auger hole, the valence band structure is more complex than the conduction band, as we saw in Subsect. 5.4.3. We must then distinguish whether this hole is excited into the light-hole band (CHLH process, Fig. 8.12b) or the spin-orbit split-off band (CHSH process, Fig. 8.12c).

In all three cases, the total energy and the total momentum (i.e., $\hbar\,\vec{k}$) of the system constituted by the three particles must be conserved.

Similar to the direct band-to-band recombination, the Auger recombination rates are expressed in units of $\mathrm{cm}^{-3}\cdot\mathrm{s}^{-1}$ and are proportional, in all three processes, to the density of electrons in the conduction band n and that of holes in the valence band p, because these are the particles which are recombining.

In the CHCC case, this rate is also proportional to the density of electrons which are susceptible to be excited, i.e., n again. The recombination rate in the CHCC process is therefore given by:

$$R_{\mathrm{CHCC}} = r_1 n^2 p \tag{8.95}$$

where r_1 is the Auger recombination coefficient for this case and is expressed in units of cm^{-1}.

For the CHLH and CHSH processes, the same argument leads to a compounded recombination rate equal to:

$$R_{\mathrm{CHLH+CHSH}} = r_2 n p^2 \tag{8.96}$$

where r_2 is the Auger recombination coefficient when Auger holes are excited.

The total Auger recombination rate is therefore:

$$R = R_{\mathrm{CHCC}} + R_{\mathrm{CHLH+CHSH}} = r_1 n^2 p + r_2 n p^2 \tag{8.97}$$

We can now follow the same analysis as the one conducted for the direct band-to-band recombination in order to determine the Auger recombination lifetime. We start from the rate Eq. (8.59). At equilibrium, $\frac{dn}{dt} = 0$ and $G = 0$, and the thermal generation rate is thus equal to:

$$G_{\mathrm{t}} = R = r_1 n_0^2 p_0 + r_2 n_0 p_0^2 \tag{8.98}$$

Let us now consider the relaxation process, which occurs after the external source of generation is removed ($G = 0$). Taking into account Eqs. (8.97) and (8.98), Eq. (8.59) becomes:

$$-\frac{d(\Delta n)}{dt} = R - G_{\mathrm{t}} = r_1\left(n^2 p - n_0^2 p_0\right) + r_2\left(n p^2 - n_0 {p_0}^2\right) \tag{8.99}$$

where $\Delta n = \Delta p$ is the excess electron and hole concentrations defined in Eq. (8.61). This expression can be expanded using Eq. (8.61), and we obtain:

$$\begin{aligned} -\frac{d(\Delta n)}{dt} &= -r_1\left[n_0^2 p_0 - (n_0 + \Delta n)^2 (p_0 + \Delta n)\right] - r_2\left[n_0 p_0^2 - (n_0 + \Delta n)(p_0 + \Delta n)^2\right] \\ &= r_1\left[\left(n_0^2 + 2n_0 p_0\right)\Delta n + (2n_0 + p_0)(\Delta n)^2 + (\Delta n)^3\right] \\ &\quad + r_2\left[\left(p_0^2 + 2n_0 p_0\right)\Delta n + (2p_0 + n_0)(\Delta n)^2 + (\Delta n)^3\right] \end{aligned}$$

We can now introduce the Auger recombination lifetime $\tau_{\mathrm{n}} = \tau_{\mathrm{p}}$ such that:

$$-\frac{d(\Delta n)}{dt} = -\frac{d(\Delta p)}{dt} = \frac{\Delta p}{\tau_{\mathrm{p}}} = \frac{\Delta n}{\tau_{\mathrm{n}}}$$

$$\tau_{\mathrm{n}}(t) = \tau_{\mathrm{p}}(t) = 1 \begin{array}{l} r_1\left[\left(n_0^2 + 2n_0 p_0\right) + (2n_0 + p_0)\Delta n + (\Delta n)^2\right] \\ + r_2\left[\left(p_0^2 + 2n_0 p_0\right) + (2p_0 + n_0)\Delta n + (\Delta n)^2\right] \end{array} \tag{8.100}$$

which becomes independent of time for weak excitation levels $\Delta n << n_0, p_0$:

$$\tau_{\mathrm{n}} = \tau_{\mathrm{p}} = \frac{1}{r_1\left(n_0^2 + 2n_0 p_0\right) + r_2\left(p_0^2 + 2n_0 p_0\right)} \tag{8.101}$$

8.6.5 Surface Recombination

The surface of a semiconductor is a violation of the crystal periodicity and therefore gives rise to energy levels near the surface which lie within the bandgap. These

correspond to surface traps. However, unlike the previously discussed carrier recombination mechanisms which occur in the bulk solid, surface recombination occurs at the surface of the solid. Moreover, the surface recombination takes place even in pure materials. Such processes play an important role in semiconductor device technology.

The energy levels introduced by the surface traps can be considered as a special case of recombination centers in Shockley-Read-Hall recombination mechanism. The same analysis as in Subsect. 8.6.4 can be conducted here for surface recombination, provided a surface density of recombination centers $(N_T)_s$ is used instead of the bulk density of centers N_T. All the other parameters would keep the same meaning.

The excess surface recombination rate is the number of electrons or holes which are recombined per unit area of the surface and per unit time. It is thus expressed in units of $cm^{-2} \cdot s^{-1}$ and can be obtained by analogy with the SRH recombination in Eq. (8.87):

$$(R - G_t)_s = v_{th}\sigma_n\sigma_p(N_T)_s \frac{(n_0 + p_0 + \Delta n)\Delta n}{\sigma_n(n_0 + n_T + \Delta n) + \sigma_p(p_0 + p_T + \Delta n)} \tag{8.102}$$

Here, Δn is the excess electron concentration near the surface considered. We can rewrite this relation as:

$$-\frac{d(\Delta n)}{dt} = (R - G_t)_s = S_n \Delta n \tag{8.103}$$

where:

$$S_n = v_{th}\sigma_n\sigma_p(N_T)_s \frac{(n_0 + p_0 + \Delta n)}{\sigma_n(n_0 + n_T + \Delta n) + \sigma_p(p_0 + p_T + \Delta n)} \tag{8.104}$$

This quantity is expressed in units of $cm \cdot s^{-1}$ and has thus the same dimension as a velocity. It is called the surface recombination velocity.

8.7 Quasi-Fermi Energy

In Sect. 7.5, we calculated the equilibrium electron concentration in the conduction band n_0 and the hole concentration in the valence band p_0 using the Fermi-Dirac distribution and arrived at Eqs. (7.18) and (7.27) in the general case and Eqs. (7.21) and (7.29) in the nondegenerate. For a given semiconductor material, these concentrations depended solely on a single parameter, the Fermi energy E_F.

Under non-equilibrium conditions, where the electron and hole concentrations in their respective bands are given by:

$$\begin{cases} n = n_0 + \Delta n \\ p = p_0 + \Delta p \end{cases} \tag{8.105}$$

the Fermi-Dirac distribution is not valid any more. However, it is convenient to maintain the mathematical formalism of the equations mentioned previously, and this is most often done for a nondegenerate semiconductor only.

Therefore, by analogy with Eq. (7.21), the non-equilibrium electron concentration in the conduction band is given by:

$$n = N_{\mathrm{c}} \exp\left(\frac{E_{\mathrm{F_n}} - E_{\mathrm{C}}}{k_{\mathrm{b}} T}\right) \tag{8.106}$$

where the quantity $E_{F_{\mathrm{n}}}$ is used instead of the Fermi energy E_{F}. This quantity is called the electron quasi-Fermi energy. Using this expression, Eqs. (7.21) and (8.105), we can write:

$$\frac{\Delta n}{n_0} = \frac{n}{n_0} - 1 = \exp\left(\frac{E_{\mathrm{F_n}} - E_{\mathrm{F}}}{k_{\mathrm{b}} T}\right) - 1 \tag{8.107}$$

Therefore, under non-equilibrium conditions, the difference between the quasi-Fermi level and the Fermi level determines the relative excess electron concentration with respect to the equilibrium concentrations.

Using this quasi-Fermi energy, it is possible to define a quasi-Fermi-Dirac distribution for electrons, which is analogous to Eq. (5.28) with E_{F} replaced by $E_{\mathrm{F_n}}$:

$$f_{e_{\mathrm{n}}}(E) = \frac{1}{\exp\left(\frac{E - E_{\mathrm{F_n}}}{k_{\mathrm{b}} T}\right) + 1} \tag{8.108}$$

A similar concept can be introduced for holes in the valence band. The hole quasi-Fermi energy $E_{F_{\mathrm{p}}}$ is defined such that:

$$p = N_{\mathrm{v}} \exp\left(\frac{E_{\mathrm{V}} - E_{\mathrm{F_p}}}{k_{\mathrm{b}} T}\right) \tag{8.109}$$

A quasi-Fermi-Dirac distribution for holes can also be defined by analogy with Eq. (7.23):

$$f_{h_{\mathrm{p}}}(E) = \frac{1}{\exp\left(\frac{E_{\mathrm{F_p}} - E}{k_{\mathrm{b}} T}\right) + 1} \tag{8.110}$$

The quasi-Fermi-Dirac distributions allow separate mathematical computations for electrons and holes in an easier manner. At equilibrium, the electron and hole quasi-Fermi energies are both equal to the Fermi energy, i.e., $E_{\mathrm{F_n}} = E_{\mathrm{F_p}} = E_{\mathrm{F}}$.

Example

Q Estimate the difference between the quasi-Fermi energies E_{Fn} and E_{Fp} and the Fermi energy E_F in an intrinsic semiconductor, given that the excess carrier concentration $\Delta n = \Delta p$ is 1% of n_0.

A The quasi-Fermi energies E_{Fn} and E_{Fp} are related to the excess carrier concentration through the expression $E_{F_n} - E_F = k_b T \ln\left(\frac{\Delta n}{n_0}\right)$ and $E_F - E_{F_p} = k_b T \ln\left(\frac{\Delta p}{p_0}\right)$, where n_0 and p_0 are the equilibrium electron and hole concentrations and are both equal to the intrinsic carrier concentration n_i since the semiconductor is assumed intrinsic at equilibrium. Therefore $\frac{\Delta n}{n_0} = \frac{\Delta p}{p_0} = 0.01$ and we obtain: $E_{F_n} - E_F = E_F - E_{F_p} = 0.0095 k_b T$.

8.8 Transport Theory: Beyond Drude

In this chapter we derived the electrical conductivity of materials using a very simple classical Newton's laws approach. We did this because the so-called Drude theory of conductivity is surprisingly powerful and useful. But it does not include the Pauli principle, for example, and does not incorporate the concept of the Fermi distribution and the Fermi level. There are many situations in which the simple Drude theory is not adequate. So we will here derive a more rigorous transport theory based on the work of Ludwig Eduard Boltzmann, and we will show how it differs from Drude, and in what special limits it reduces to the Drude theory.

8.8.1 The Boltzmann Equation

We must start with the concept of the distribution function of electrons $f_{\vec{k}}(\vec{r}, t)$. This quantity is the probability of an electron occupying the Bloch state $\boldsymbol{k}$ in the solid at position r at time t. Note that this is not the equilibrium Fermi distribution function $f^{\,0}_{\vec{k}}$ from Eq. (5.28) which only depends on the energy and Fermi energy. The new non-equilibrium distribution $f_{\vec{k}}(\vec{r}, t)$ tells us how many particles there are in this region of space at time t and with momentum k. In a steady-state situation, the total rate of change with time of the distribution function must be zero. Specifically, there are changes in the function $f_{\vec{k}}(\vec{r}, t)$ caused by specific processes. Thus the distribution changes because:

1. The particles in the material are diffusing in space.
2. Electric and magnetic fields are accelerating the particles.
3. There are scattering processes which change the momentum and energies of the particles. These processes include scattering from impurities, defects, phonons, etc., all processes which break the Bloch symmetry of the crystal.

The information we seek is in $f_{\vec{k}}(\vec{r},t)$. Knowing this function we can compute the current via

$$\vec{J} = -\int d\vec{k}\, q v_{\vec{k}} f(\vec{k},\vec{r},t) \tag{8.111}$$

where $v_{\vec{k}}$ is the velocity. To calculate the distribution function, we now examine each of the above processes in turn. First we note that because the particles diffuse in space, one source of time variation is "diffusion" which is described by the variation:

$$\left.\frac{\partial f}{\partial t}\right|_{\text{diffusion}} = -\vec{v}_{\vec{k}}.\nabla f_{\vec{k}} \rightarrow \frac{\partial f}{\partial t} = \frac{\partial \vec{r}}{\partial t}.\frac{\partial f_{\vec{k}}}{\partial \vec{r}} \tag{8.112}$$

Then there is the influence of external fields. To proceed we remember from Eqs. (8.1) and (8.14) that with Bloch states, the Newton laws act on the pseudo momentum parameter k as:

$$\frac{d\vec{k}}{dt} = -\frac{q}{\hbar}\left(\vec{E}_{\text{a}} + \vec{v}_{\vec{k}} \times \vec{B}\right) \rightarrow \text{MKS} \tag{8.113}$$

where E_a is the applied field. Thus it follows that the field variation is:

$$\left.\frac{\partial f}{\partial t}\right|_{\text{field}} = -\frac{d\vec{k}}{dt}.\frac{\partial f(\vec{k},\vec{r},t)}{\partial \vec{k}} = \frac{q}{\hbar}\left(\vec{E}_{\text{a}} + \vec{v} \times \vec{B}\right).\frac{\partial f(\vec{k},\vec{r},t)}{\partial \vec{k}} \tag{8.114}$$

$$v_{\vec{k}} = \frac{1}{\hbar}\vec{\nabla} E_{\vec{k}} \tag{8.115}$$

Finally the change due to collisions can be written in terms of a generalized rate equation:

$$\left.\frac{\partial f}{\partial t}\right|_{\text{collisons}} = \int \left\{f_{\vec{k}}\left(1 - f_{\vec{k}'}\right) - f_{\vec{k}'}\left(1 - f_{\vec{k}}\right)\right\} W\left(\vec{k},\vec{k}'\right) d\vec{k}' \tag{8.116}$$

where $W\left(\vec{k},\vec{k}'\right)$ is the rate at which electrons are scattered from k to k'.

In the steady state, the sum of all variations must add up to zero:

$$\left.\frac{\partial f_{\vec{k}}}{\partial t}\right|_{\text{field}} + \left.\frac{\partial f_{\vec{k}}}{\partial t}\right|_{\text{diffusion}} + \left.\frac{\partial f_{\vec{k}}}{\partial t}\right|_{\text{collisions}} = \frac{df_{\vec{k}}}{dt} = 0 \tag{8.117}$$

So we have:

$$\begin{aligned}\frac{\partial f}{\partial t} &= \int \left\{f_{\vec{k}}\left(1-f_{\vec{k'}}\right) - f_{\vec{k'}}\left(1-f_{\vec{k}}\right)\right\} W\left(\vec{k},\vec{k'}\right) d\,\vec{k'} \\ &+\frac{q}{\hbar}\left(\vec{E}_{\mathrm{a}} + \vec{v}\times\vec{B}\right).\frac{\partial f\left(\vec{k},\vec{r},t\right)}{\partial\,\vec{k}} - \vec{v}_{\vec{k}}.\,\vec{\nabla} f_{\vec{k}} = 0\end{aligned} \tag{8.118}$$

In principle if one knows the scattering rates, then one can compute the result by following the trajectory of the particles in space and time. The Boltzmann equation can also be solved numerically using the Monte Carlo Method. Let us consider the simple first-order solution in an electric field using the relaxation time approximation. So one says the following: the field produces a small change in the distribution function which we call:

$$f = f_0 + f_1\left(\vec{k},\vec{r},t\right) \tag{8.119}$$

This deviation from the steady state must return to zero when the system has had time to relax or reach a steady state, so we can write:

$$-\frac{\partial f_k}{\partial t} = \frac{1}{\tau} f_1\left(\vec{k},\vec{r},t\right) \tag{8.120}$$

$$f_1 = f_1\left(\vec{k},\vec{r},0\right)e^{-t/\tau} \tag{8.121}$$

It is also useful to write:

$$\frac{\partial f}{\partial\,\vec{k}} = \frac{\partial E_{\vec{k}}}{\partial\,\vec{k}}\frac{\partial f}{\partial E_{\vec{k}}} \tag{8.122}$$

Substituting back into the Boltzmann equation and including only the electric field term, we have:

$$-q\frac{df}{dE_{\vec{k}}}\vec{v}_{\vec{k}}.\vec{E}_{\mathrm{a}} = -\frac{f_1}{\tau} \tag{8.123}$$

$$f_1 = q\frac{df}{dE_{\vec{k}}}\left(\vec{v}_{\vec{k}}.\vec{E}_{\mathrm{a}}\right)\tau \tag{8.124}$$

This actually already gives us the first-order solution in an electric field.

Note that in Drude theory we say that in the steady state, the field force balances the frictional force. Here it is the population which reaches a steady state, not the individual particles. Another way of looking at it is to go back to Eq. (8.119) and allow the electron momentum to be increased by the field up to a relaxation time τ, after which it is interrupted and has to start accelerating again. Thus in the steady state, $\delta k_x = -q\tau E_{ax}/\hbar$ which gives a concomitant change in the energy to $E(k_y, k_z, k_x - qE_{ax}\tau/\hbar)$. The change in the energy gives rise to a steady-state change in the distribution function f as in Eq. (8.119), E_a is the applied field, and:

$$f_{\vec{k}} = f^{0}_{\vec{k}} + \frac{\partial f^{0}_{k}(E_{\vec{k}})}{\partial E_{\vec{k}}} \frac{\partial E(\vec{k})}{\partial \vec{k}} \cdot \frac{q\tau}{\hbar} \vec{E}_{a} \tag{8.125}$$

Substituting Eq. (8.125) back into the expression for the current, we have:

$$\vec{J} = -\frac{1}{4\pi^3} \int d\vec{k}\, q v_{\vec{k}} f(\vec{k}, \vec{r}, t) = -\frac{1}{4\pi^3} \int d\vec{k}\, q v_{\vec{k}} f_1(\vec{k}, \vec{r}, t) \tag{8.126}$$

$$\vec{J} = \frac{1}{4\pi^3} q^2 \int d\vec{k}\, \tau(E_{\vec{k}}) \vec{v}_{\vec{k}} (\vec{v}_{\vec{k}} . \vec{E}_{a}) \left(-\frac{\partial f^{0}_{\vec{k}}}{\partial E_{\vec{k}}} \right) \tag{8.127}$$

Write the volume integral as an integral over energy and surface of constant energy:

$$\begin{aligned} d\vec{k} &= k^2 dk \sin(\vartheta) d\theta d\phi \\ k^2 &= 2mE/\hbar^2 \\ d\vec{k} &= m^{3/2} E^{1/2} dE \sin(\vartheta) d\theta d\phi = dE dS_F / \hbar v_{\vec{k}} \end{aligned} \tag{8.128}$$

So that:

$$J = \frac{1}{4\pi^2} \int d\vec{k}\, q^2 \tau(E_{\vec{k}}) \vec{v}_{\vec{k}} (\vec{v}_{\vec{k}} \vec{E}_{a}) \left(-\frac{\partial f^{0}_{\vec{k}}}{\partial E_{\vec{k}}} \right) \tag{8.129}$$

$$J = 2q^2 \int dE_k g_V(E_k) \tau(E_k) \vec{v}_{\vec{k}} (\vec{v}_{\vec{k}} \vec{E}_{a}) \left(-\frac{\partial f^{0}_{\vec{k}}}{\partial E_k} \right) \tag{8.130}$$

$$E_{KE} = 1/2\, mv^2 \tag{8.131}$$

$$v_{\vec{k}} v_{\vec{k}} \vec{E}_{a} \rightarrow v_x^2 E_{a,x} = \frac{2E}{3m} E_{a,x} \tag{8.132}$$

where in order to avoid confusion, we use E_a for applied field and we use g_V to denote the density of states per unit volume. With the low-temperature form:

$$-\frac{\partial f}{\partial E} = \delta(E - E_F) \tag{8.133}$$

The current density reduces to the form (where $\bar{d}$ is the dimensionality of the system):

$$J = \frac{q^2\tau(E_F)}{\bar{d}} v_F^2 g_V(E_F) E_{a,x} \tag{8.134}$$

8.8.2 Connection to Drude Theory

In order to relate this expression to the familiar Drude result from Eq. (8.7), we consider three dimensions and also assume that the relaxation time is energy independent and that we are dealing with nearly free electrons:

$$J = \frac{q^2\tau}{3} v_F^2 g(E_F) E_{a,x} \sim \frac{1}{3} m^* v_F^2 g(E_F) \frac{q^2\tau}{m^*} \tag{8.135}$$

$$J = n\frac{q^2\tau}{m^*} E_{a,x} \rightarrow n = \frac{1}{3} m^* v_F^2 g(E_F) \tag{8.136}$$

$$\sigma = \frac{q^2\tau(E_F)}{3} v_F^2 g(E_F) \rightarrow n\frac{q^2\tau}{m^*} \tag{8.137}$$

Here n is the effective carrier density.

The connection is made and we see why Drude represents a serious approximation:

(i) The more correct i.e, Boltzmann form of the conductivity scales with the density of states at the Fermi level. So that if there is no free charge that responds to an applied field, there is no conduction.
(ii) The Boltzmann equation includes the Fermi distribution, so the Pauli principle is obeyed.
(iii) The Boltzmann equation result allows the relaxation time to be energy dependent. This energy dependence enters the result via the integral Eq. (8.130) which then also takes care of the temperature dependence.
(iv) The Boltzmann equation result allows the group velocity to deviate from the nearly free electron law.

In general we see therefore that the Boltzmann equation conductivity is far superior to the Drude theory and is really the right way to proceed in a Bloch solid.

8.9 Summary

In this chapter, we have covered a few important non-equilibrium transport phenomena involving charge carriers. Firstly, we discussed the electrical conductivity (Ohm's law) in the presence of an external electric field. There, we introduced the concepts of conductivity, resistivity, as well as carrier collision or scattering. Then, secondly we described the Hall effect for an *n*-type and then a *p*-type semiconductor in the presence of perpendicular electric and magnetic fields. There, we introduced the notion of carrier mobility. Thirdly, we discussed the diffusion of charge carriers in an inhomogeneous semiconductor, leading to the concepts of diffusion length and the Einstein relations.

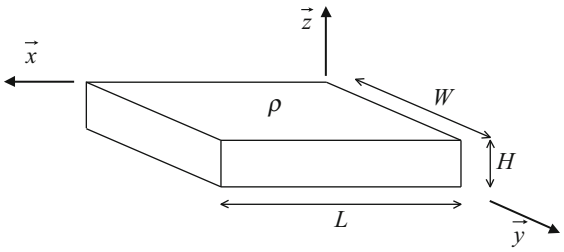
The recombination mechanisms of charge carriers in a semiconductor have been described, including the direct band-to-band, Shockley-Read-Hall, Auger, and surface recombination processes. The concepts of recombination lifetime and capture cross section were introduced.

We introduced the notion of quasi-Fermi energy to describe the electron and hole distribution under non-equilibrium conditions while at the same time maintaining the same mathematical formalism as under equilibrium conditions.

In the last part of the chapter, we introduced the reader to a more powerful transport description known as the Boltzmann equation approach. We derived a more general formula for the conductivity, and we showed why it is superior to the Drude method.

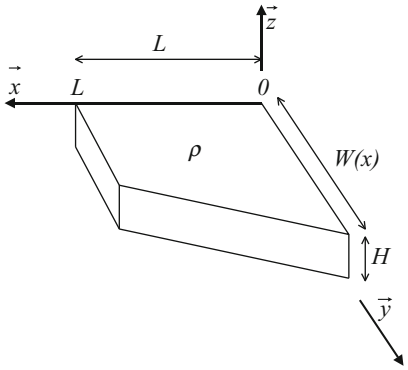
Problems

1. Consider the semiconductor slab shown in the figure below with dimensions $L = 1$ cm, $W = 0.2$ cm, and $H = 0.25$ cm and with a resistivity of 0.01 Ω·cm. What would be the resistance one would measure across opposite faces in all three directions (x, y, and z)? Knowing there is a uniform concentration $n = 10^{16}$ cm^{-3} of electrons in this semiconductor (and no holes), calculate the mobility of these electrons.



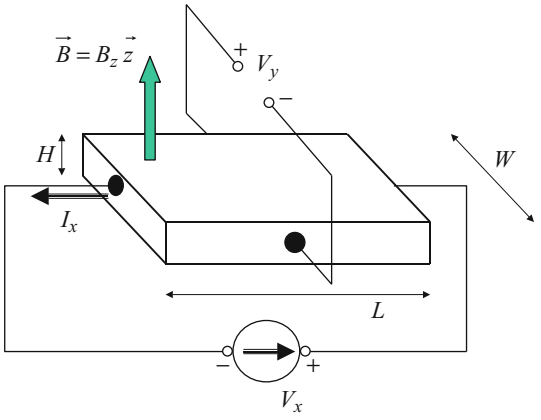
2. Consider the semiconductor block with a resistivity of 0.01 Ω·cm as shown in the figure below. The width of this block is constant but follows the relation $W = 1 + 2(L - x)$ cm when x is varied from 0 to L. The other dimensions are

$L = 1$ cm and $H = 0.25$ cm. Calculate the resistance in the x-direction. For this, you may consider the semiconductor block as a series of parallelepiped slabs next to one another.

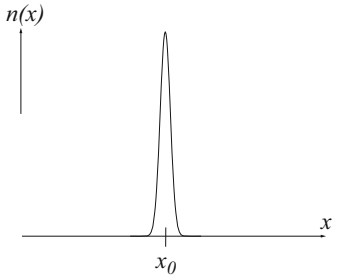


3. Do the same as in Problem 2, but in the y-direction.
4. Consider the Hall effect measurement experiment depicted in the figure below. The dimensions of the semiconductor slab are $L = 2$ mm, $W = 1$ mm, and $H = 2$ µm. Assume the current $I_x = 10$ mA, the voltages $V_x = 10$ V and $V_y = -4$ V, and a magnetic induction $B_z = 0.05$ T.

   Determine if the semiconductor is n-type or p-type, the Hall constant, the carrier concentration, the Hall mobility, the conductivity, and the resistivity of the semiconductor (assumed uniform).

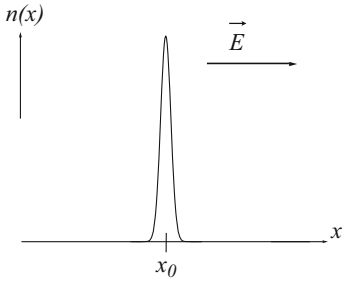


5. Consider an experiment where excess electrons are generated in a "burst" at $t = 0$ and $x = x_0$ in a semiconductor, resulting in the concentration profile $n(x)$ shown in the figure below.



Draw the shape of the concentration profile $n(x)$ as a result of the one-dimensional diffusion in the x-direction. No other external forces are present. Draw several shapes corresponding to several times after the initial "burst."

6. Do the same as in Problem 5, but consider, in addition, that there is an electric field strength $\vec{E}$ in the direction as shown in the figure below.



7. The electron mobility in a Ge crystal is experimentally found to be proportional to $T^{-1.66}$ (i.e., the mobility decreases with increasing temperature). Knowing that this mobility is 4000 cm^2/Vs at 300 K, determine the electron diffusion coefficient at 300 K and 77 K. Compare.
8. Consider an n-type Si semiconductor at room temperature with an excess electron concentration which decreases from 4×10^{16} cm^{-3} to 1 cm^{-3} (practically zero) over a distance of 1 mm. Determine the diffusion length of these electrons.
9. Assume a one-dimensional model in which holes are generated at a rate of $G(x,t)$. Let τ_p be the recombination lifetime for holes and p_0 be the equilibrium hole concentration. Give an expression for $\frac{\partial p(x,t)}{\partial t}$, i.e., the rate of change for the hole concentration at position x, as a function of the diffusion current $J_h^{diff}(x,t)$ and the parameters defined previously. This relation is called a continuity equation and states that the total number of holes must be accounted for. Using Eq. (8.42), rewrite this relation such that it involves the hole concentration $p(x,t)$ as the only unknown.

Further Reading

Anselm A (1981) Introduction to semiconductor theory. Prentice-Hall, Englewood Cliffs, NJ
Ashcroft NW, Mermin ND (1976) *Solid State Physics*, Holt. Rinehart and Winston, New York
Cohen MM (1972) Introduction to the quantum theory of semiconductors. Gordon and Breach, New York
Esaki L (ed) (1991) Highlights in condensed matter physics and future concepts, vol 285. Plenum Press, New YorkNATO Science Forum Series
Ferry DK (1991) Semiconductors. Macmillan, New York
Hummel RE (1986) Electronic properties of materials. Springer-Verlag, New York
Orton JW, Blood P (1990) The electrical characterization of semiconductors: measurement of minority carrier properties. Academic Press, San Diego
Pankove JI (1975) Optical processes in semiconductors. Dover, New York
Peyghambarian N, Koch SW, Mysyrowicz A (1993) Introduction to semiconductor optics. Prentice-Hall, Englewood Cliffs, NJ
Pierret RF (1989) Advanced semiconductor fundamentals. Addison-Wesley, Reading, MA
Pollock CR (1995) Fundamentals of optoelectronics. Irwin, Burr Ridge, IL
Ridley BK (1999) Quantum processes in semiconductors. Oxford University Press, New York
Rogalski A (1995) Infrared photon detectors, Bellingham, Washington
Streetman BG (1990) Solid state electronic devices. Prentice-Hall, Englewood Cliffs, NJ
Sze SM (1981) Physics of semiconductor devices. John Wiley & Sons, New York
Wang S (1989) Fundamentals of semiconductor theory and device physics. Prentice-Hall, Englewood Cliffs, NJ
Wolfe CM, Holonyak N Jr, Stillman GE (1989) Physical properties of semiconductors. Prentice-Hall, Englewood Cliffs, NJ

Semiconductor *p-n* and Metal-Semiconductor Junctions

9

9.1 Introduction

Until now, our discussion was based solely on homogeneous semiconductors whose properties are uniform in space. Although a few devices can be made from such semiconductors, the majority of devices and the most important ones utilize nonhomogeneous semiconductor structures. Most of them involve semiconductor *p-n* junctions, in which a *p*-type doped region and an *n*-type doped region are brought into contact. Such a junction actually forms an electrical diode. This is why it is usual to talk about a *p-n* junction as a diode. Another important structure involves a semiconductor in intimate contact with a metal, leading to what is called a metal-semiconductor junction. Under certain circumstances, this configuration can also lead to an electrical diode.

The objective of this chapter will first be to establish an accurate model for the *p-n* junction which can be at the same time mathematically described. This model will be the ideal p-n junction diode. The basic properties of this ideal p-n junction at equilibrium will be described in detail. The non-equilibrium properties of this *p-n* junction will then be discussed by deriving the diode equation which relates the current and voltage across the diode. Deviations from the ideal diode case will also be described. Finally, this chapter will also discuss the properties of metal-semiconductor junctions and compare them with those of *p-n* junctions.

9.2 Ideal *p-n* Junction at Equilibrium

9.2.1 Ideal *p-n* Junction

The ideal *p-n* junction model is also called the abrupt junction or step junction model. This is an idealized model for which we assume that the material is uniformly doped *p*-type with a total acceptor concentration N_A on one side of the junction (e.g.,

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_9

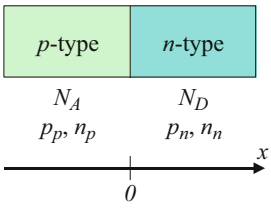


Fig. 9.1 Ideal *p-n* junction model, in which one side of the junction is a purely *p*-type semiconductor and the other a purely *n*-type semiconductor. Both materials are uniformly doped

$x < 0$), and the material is uniformly doped *n*-type with a total donor concentration N_D on the other side (e.g., $x > 0$). For further simplicity, we will consider a homojunction, i.e., both doped regions are of the same semiconductor material. We will restrict our analysis to the one-dimensional case, as illustrated in Fig. 9.1.

In the *p*-type doped region far from the junction area, the equilibrium hole and electron concentrations are denoted p_p and n_p, respectively. In the *n*-type doped region far from the junction area, the hole and electron concentrations are denoted p_n and n_n, respectively. These carrier concentrations satisfy the mass action law in Eq. (7.31):

$$p_p n_p = p_n n_n = n_i^2 \tag{9.1}$$

where n_i is the intrinsic carrier concentration in the semiconductor material considered. We further assume that all the dopants are ionized, which leads to the following carrier concentrations for the *p*- and *n*-type regions, respectively:

$$\begin{cases} p_p = N_A \left(10^{16}\text{cm}^{-3}\right) \\ n_p = \dfrac{n_i^2}{N_A}\left(10^{5}\text{cm}^{-3}\right) \end{cases} \text{and} \begin{cases} n_n = N_D \left(10^{17}\text{cm}^{-3}\right) \\ p_n = \dfrac{n_i^2}{N_D}\left(10^{4}\text{cm}^{-3}\right) \end{cases} \tag{9.2}$$

A few typical values for these concentrations are given in parenthesis. It is important to remember that both a *p*-type, and an *n*-type, isolated semiconductors are electrically neutral.

9.2.2 Depletion Approximation

However, when bringing a *p*-type semiconductor into contact with an *n*-type semiconductor, the material is not electrically neutral everywhere anymore. Indeed, on one side of the junction area, for $x < 0$, there is a high concentration of holes, whereas on the other side there is a low concentration of holes. This asymmetry in carrier density results in the diffusion of holes across the junction as shown in Fig. 9.1. By doing so, the holes leave behind uncompensated acceptors ($x < 0$) which are negatively charged. A similar analysis can be carried out for electrons as there is also an asymmetry in the density of electrons on either side of the *p-n* junction. This leads to their diffusion and makes the material positively charged for $x > 0$ as the electrons leave behind uncompensated donors, as shown in Fig. 9.2.

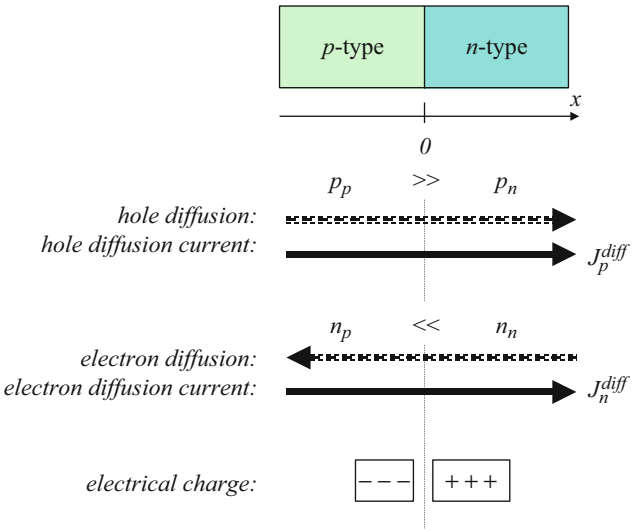


Fig. 9.2 Hole and electron diffusion across a *p-n* junction. The holes diffuse from the left to the right, which leads to a diffusion electrical current from the left to the right as well. By contrast, the electrons diffuse from the right to the left, but this leads to a diffusion electrical current from the left to the right because of the negative charge of electrons. The diffusion process leaves uncompensated acceptors in the *p*-type region and donors in the *n*-type regions, i.e., a net negative charge in the *p*-type region and a net positive charge in the *n*-type region. The presence of these charges results in a built-in electric field

This redistribution of electrical charge does not endure indefinitely. Indeed, as positive and negative charges appear on the $x > 0$ and $x < 0$ sides of the junction, respectively, an electric field strength $E(x)$, called the built-in electric field, will result and is shown in Fig. 9.3 As discussed in Chap. 8, this electric field will generate the drift of the positively charged holes and the negatively charged electrons. By comparing Figs. 9.2 and 9.3, we can see that the drift of these charge carriers counteracts the previous diffusion process. An equilibrium state is reached when the diffusion currents $J^{\text{diffusion}}$ and drift currents J^{drift} are exactly balanced for each type of carrier, i.e., holes and electrons taken independently:

$$\begin{cases} J_{\text{h}}^{\text{diff}} + J_{\text{h}}^{\text{drift}} = 0 \\ J_{\text{e}}^{\text{diff}} + J_{\text{e}}^{\text{drift}} = 0 \end{cases} \tag{9.3}$$

There is a transition region around the *p-n* junction area with a width W_0 in which the electrical charges are present. This region is called the space charge region and is schematically shown in Fig. 9.4a. The charge distribution within this region is modeled as follows: we consider that there is a uniform concentration of negative charges for $-x_{\text{p0}} < x < 0$ equal to $Q(x) = -qN_{\text{A}}$ (where N_{A} is the total concentration of acceptors in the *p*-type region) and a uniform concentration of positive charges for $0 < x < x_{\text{n0}}$ and equal to $Q(x) = +qN_{\text{D}}$ (where N_{D} is the total concentration of donors

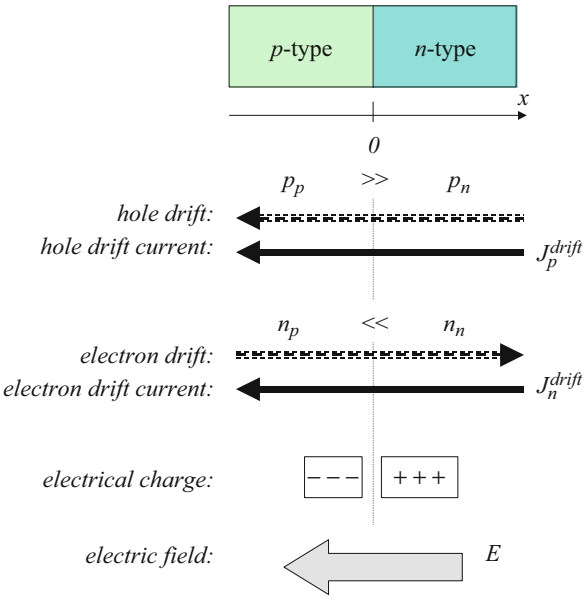


Fig. 9.3 Hole and electron drift across a *p-n* junction. Under the influence of the built-in electric field, the holes drift from the right to the left, which leads to a drift electrical current from the right to the left as well. By contrast, the electrons drift from the left to the right, but this leads to a drift electrical current from the right to the left because of the negative charge of electrons. The drift process counterbalances the diffusion of charge carriers in order to bring the system into equilibrium

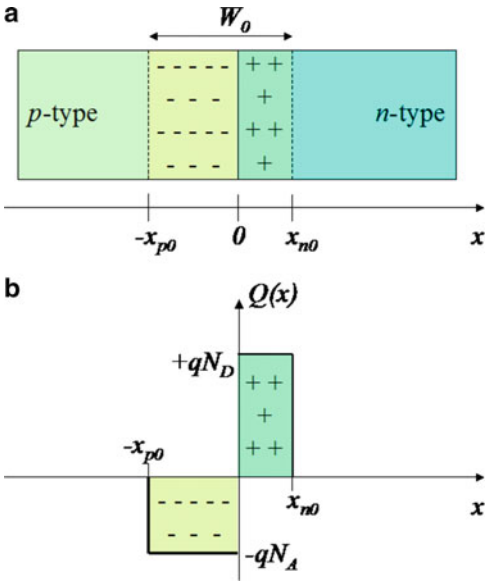


Fig. 9.4 (**a**) Space charge region in a *p-n* junction. Near the junction area, the *p*-type region is negatively charged as a result of the diffusion of charge carriers. (**b**) Electrical charge density in a *p-n* junction. To keep the overall charge neutrality, the total number of negative charges in the *p*-type region is equal to the total number of positive charges in the *n*-type region. In the depletion approximation, the charges are assumed uniformly distributed in space, within the depletion region delimited by $-x_{p0}$ and x_{n0}

in the *n*-type region). The quantities x_{p0} and x_{n0} are positive and express how much the space charge region extends on each side of the junction, as illustrated in Fig. 9.4b. The width of the space charge region, also called depletion width, is then given by:

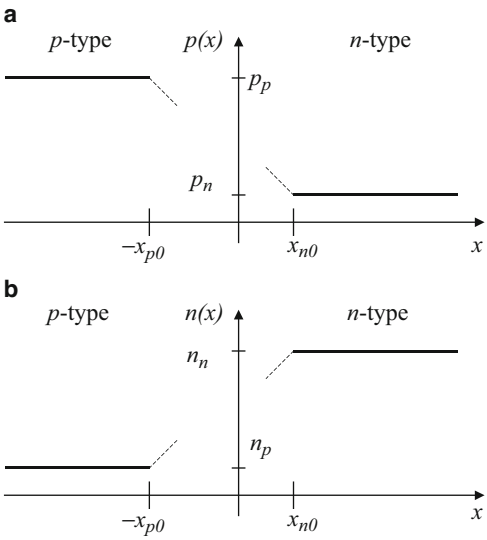


Fig. 9.5 (**a**) Hole and (**b**) electron concentrations in a *p-n* junction. In the depletion approximation, the hole and electron concentrations are assumed to be constant and equal to their equilibrium values outside of the depletion region

$$W_0 = x_{\mathrm{n0}} + x_{\mathrm{p0}} \tag{9.4}$$

Outside of this space charge region, we assume that the semiconductor is electrically neutral without any charge depletion and that the hole and electron concentrations are given by Eq. (9.2). These regions will be called the bulk *p*-type and bulk *n*-type region. The carrier concentrations must therefore somehow go from a high value on one side of the junction to a low value on the other side, and this occurs within the space charge region, as illustrated in Fig. 9.4 In particular, we have (Fig. 9.5):

$$\begin{cases} p\left(-x_{\mathrm{p0}}\right) = p_{\mathrm{p}} & \text{and} \quad p(x_{\mathrm{n0}}) = p_{\mathrm{n}} \\ n\left(-x_{\mathrm{p0}}\right) = n_{\mathrm{p}} & \text{and} \quad n(x_{\mathrm{n0}}) = n_{\mathrm{n}} \end{cases} \tag{9.5}$$

This model is called the depletion approximation. In this model, there are no free holes or electrons in the space charge region: the depletion of carriers is complete. The electric field exists only within this space charge region.

Because the entire *p-n* structure must globally remain electrically neutral, and therefore the space charge region must be neutral as a whole, we must equate the total number of negative charges on one side of the junction to the total number of positive charges on the other side, i.e.:

$$qAN_{\mathrm{A}}x_{\mathrm{p0}} = qAN_{\mathrm{D}}x_{\mathrm{n0}}$$

where A is the cross-section area of the junction, and after simplification:

$$N_{\mathrm{A}}x_{\mathrm{p0}} = N_{\mathrm{D}}x_{\mathrm{n0}} \tag{9.6}$$

Combining Eqs. (9.4) and (9.6), we can express the quantities x_{p0} and x_{n0} as a function of the depletion width W_0:

$$\begin{cases} x_{p0} = \dfrac{N_D}{N_A + N_D} W_0 \\ x_{n0} = \dfrac{N_A}{N_A + N_D} W_0 \end{cases} \tag{9.7}$$

These show that the space charge region extends more in the *p*-type region than in the *n*-type region when $N_D > N_A$ and reciprocally.

Example

Q Estimate the thickness ratio of the depletion region in the *p*-type side ($N_A = 10^{18}$ cm^{-3}) and the *n*-type side ($N_D = 10^{17}$ cm^{-3}) for an abrupt *p-n* junction in the depletion approximation.

A The thicknesses of the depletion region in the *p*-type side and the *n*-type side are denoted x_{p0} and x_{n0}, respectively. Their ratio is such that:

$$\frac{x_{p0}}{x_{n0}} = \frac{N_D}{N_A} = \frac{10^{17}}{10^{18}} = 0.1 \cdot$$

9.2.3 Built-In Electric Field

The built-in electric field strength can be calculated using Gauss's law which can be written in our one-dimensional model as:

$$\frac{dE(x)}{dx} = \frac{Q(x)}{\varepsilon} \tag{9.8}$$

where ε is the permittivity of the semiconductor material and $Q(x)$ is the total charge concentration. This relation can be rewritten for either sides of the junction:

$$\begin{cases} \dfrac{dE(x)}{dx} = -\dfrac{qN_A}{\varepsilon} & \text{for} - x_{p0} < x < 0 \\ \dfrac{dE(x)}{dx} = \dfrac{qN_E}{\varepsilon} & \text{for} \, 0 < x < x_{n0} \end{cases} \tag{9.9}$$

From these relations we see that the electric field strength varies linearly on either side of the junction. By integrating Eq. (9.8) using the boundary conditions assumed in the depletion approximation:

$$E(-x_{p0}) = E(x_{n0}) = 0 \tag{9.10}$$

that the electric field strength is equal to zero at the limits of the space charge region ($x = -x_{p0}$ and $x = x_{n0}$), we obtain successively:

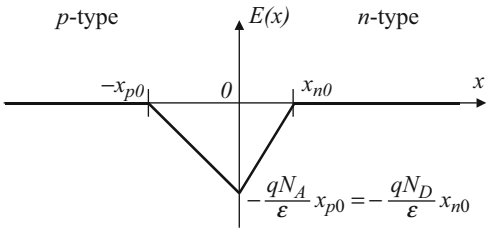


Fig. 9.6 Built-in electric field strength profile across a *p-n* junction. In the depletion approximation, the electric field strength is zero outside the depletion region because there is no net electrical charge. Within the depletion region, the electric field strength varies linearly with distance

$$\begin{cases} E(x) = \int_{-x_{p0}}^{x} dEdx = \int_{-x_{p0}}^{x} -\frac{qN_A}{\varepsilon}dx & \text{for } -x_{p0} < x < 0 \\ E(x) = \int_{x_{n0}}^{x} dEdx = \int_{x_{n0}}^{x} \frac{qN_D}{\varepsilon}dx & \text{for } 0 < x < x_{n0} \end{cases}$$

$$\begin{cases} E(x) = -\frac{qN_A}{\varepsilon}(x + x_{p0}) & \text{for } -x_{p0} < x < 0 \\ E(x) = \frac{qN_D}{\varepsilon}(x - x_{n0}) & \text{for } 0 < x < x_{n0} \end{cases} \tag{9.11}$$

For $x = 0$, we obtain two expressions for the electric field strength from the two previous expressions for $E(x)$:

$$\begin{cases} E(0) = -\frac{qN_A}{\varepsilon}(x_{p0}) \\ E(0) = \frac{qN_D}{\varepsilon}(-x_{n0}) \end{cases} \tag{9.12}$$

And these expressions are equal, according to Eq. (9.6). Therefore, the global electrical neutrality of the *p-n* structure ensures the continuity of the built-in electric field strength. A plot of $E(x)$ is shown in Fig. 9.6.

9.2.4 Built-In Potential

As a result of the presence of an electric field, an electrical potential $V(x)$ also exists and is related to the electric field strength through:

$$E(x) = -\frac{dV(x)}{dx} \tag{9.13}$$

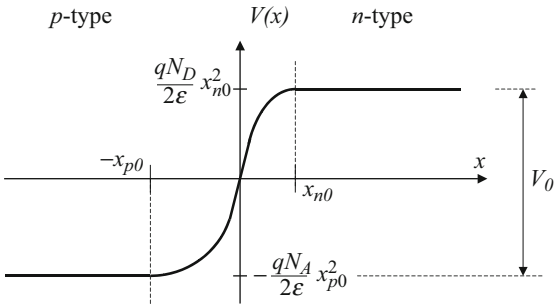


Fig. 9.7 Built-in potential profile across a *p-n* junction. In the depletion approximation, there is no variation of the potential outside the depletion region

The potential is constant outside the space charge region because the electric field strength is equal to zero there. An analytical expression for the electrical potential can be obtained by integrating Eq. (9.11):

$$\begin{cases} V(x) = \dfrac{qN_{\mathrm{A}}}{\varepsilon}\left(\dfrac{x^2}{2} + x_{\mathrm{p0}}x\right) & \text{for} - x_{\mathrm{p0}} < x < 0 \\ V(x) = -\dfrac{qN_{\mathrm{D}}}{\varepsilon}\left(\dfrac{x^2}{2} - x_{\mathrm{n0}}x\right) & \text{for } 0 < x < x_{\mathrm{n0}} \end{cases} \tag{9.14}$$

where we chose the origin of the potential at $x = 0$ and applied the continuity condition of the potential at $x = 0$. This potential is plotted in Figs. 9.6 and 9.7.

The total potential difference across the *p-n* junction is called the built-in potential and is conventionally denoted V_{bi} or V_0. It can be obtained by evaluating the potential difference between $x = -x_{\mathrm{p0}}$ and $x = x_{\mathrm{n0}}$:

$$V_0 = V(x_{\mathrm{n0}}) - V\left(-x_{\mathrm{p0}}\right) \tag{9.15}$$

This can be rewritten as:

$$V_0 = \frac{qN_{\mathrm{D}}}{\varepsilon}\frac{x_{\mathrm{n}}^2 0}{2} + \frac{qN_{\mathrm{A}}}{\varepsilon}\frac{x_{\mathrm{p}}^2 0}{2} \tag{9.16}$$

Expressing $-x_{\mathrm{p0}}$ and x_{n0} as a function of the depletion width given in Eq. (9.7), we obtain:

$$V_0 = \frac{q}{2\varepsilon}\frac{N_{\mathrm{A}}N_{\mathrm{D}}}{(N_{\mathrm{A}} + N_{\mathrm{D}})}W_0^2 \tag{9.17}$$

Another independent expression of the built-in potential can be obtained by expressing the balancing of the diffusion and drift currents. In Chap. 8 we determined analytical expressions for these currents in Eqs. (8.12) and (8.38) for holes and Eqs. (8.12) and (8.36) for electrons. The total current from the motion of holes and that from the motion of electrons are given by:

$$\begin{cases} J_{\mathrm{h}}^{\mathrm{diff}} + J_{\mathrm{h}}^{\mathrm{drift}} = -qD_{\mathrm{p}} \dfrac{dp(x)}{dx} + q\mu_{\mathrm{h}} p(x) E(x) \\ J_{\mathrm{e}}^{\mathrm{diff}} + J_{\mathrm{e}}^{\mathrm{drift}} = qD_{\mathrm{n}} \dfrac{dn(x)}{dx} + q\mu_{\mathrm{e}} n(x) E(x) \end{cases} \tag{9.18}$$

In these expressions, $p(x)$ and $n(x)$ represent the hole and electron concentrations at a position x. Taking into account the condition of Eq. (9.3) stating the exact balancing of the diffusion and drift currents for holes and electrons, we can write:

$$\begin{cases} D_{\mathrm{p}} \dfrac{dp(x)}{dx} = \mu_{\mathrm{h}} p(x) E(x) \\ D_{\mathrm{n}} \dfrac{dn(x)}{dx} = -\mu_{\mathrm{e}} n(e) E(x) \end{cases} \tag{9.19}$$

which can be rewritten using Eq. (9.19) as:

$$\begin{cases} \dfrac{D_{\mathrm{p}}}{\mu_{\mathrm{h}}} \dfrac{1}{p(x)} \dfrac{dp(x)}{dx} = -\dfrac{dV(x)}{dx} \\ \dfrac{D_{\mathrm{n}}}{\mu_{\mathrm{e}}} \dfrac{1}{n(x)} \dfrac{dn(x)}{dx} = \dfrac{dV(x)}{dx} \end{cases}$$

By integrating these equations, we get successively:

$$\begin{cases} \dfrac{D_{\mathrm{p}}}{\mu_{\mathrm{h}}} \displaystyle\int_{-x_{\mathrm{p0}}}^{x_{\mathrm{n0}}} \frac{1}{p(x)} \frac{dp(x)}{dx} dx = -\int_{-x_{\mathrm{p0}}}^{x_{\mathrm{n0}}} \frac{dV(x)}{dx} dx \\ \dfrac{D_{\mathrm{n}}}{\mu_{\mathrm{e}}} \displaystyle\int_{-x_{\mathrm{p0}}}^{x_{\mathrm{n0}}} \frac{1}{n(x)} \frac{dn(x)}{dx} dx = \int_{-x_{\mathrm{p0}}}^{x_{\mathrm{n0}}} \frac{dV(x)}{dx} dx \end{cases}$$

Using Eqs. (9.5) and (9.15), and by taking into account the Einstein relations $\frac{D_{\mathrm{p}}}{\mu_{\mathrm{h}}} = \frac{D_{\mathrm{n}}}{\mu_{\mathrm{e}}} = \frac{k_{\mathrm{b}}T}{q}$ obtained from Eqs. (8.44) and (8.45), we get:

$$\begin{cases} \dfrac{k_{\mathrm{b}}T}{q} \displaystyle\int_{p_{\mathrm{p}}}^{p_{\mathrm{n}}} \frac{dp}{p} = -\int_{0}^{V_0} dV \\ \dfrac{k_{\mathrm{b}}T}{q} \displaystyle\int_{n_{\mathrm{p}}}^{n_{\mathrm{n}}} \frac{dn}{n} = \int_{0}^{V_0} dV \end{cases}$$

which integrates easily into:

$$\begin{cases} \dfrac{k_b T}{q} \ln\left(\dfrac{p_n}{p_p}\right) = -V_0 \\ \dfrac{k_b T}{q} \ln\left(\dfrac{n_n}{n_p}\right) = V_0 \end{cases}$$

i.e.:

$$V_0 = \frac{k_b T}{q} \ln\left(\frac{p_p}{p_n}\right) = \frac{k_b T}{q} \ln\left(\frac{n_n}{n_p}\right) \tag{9.20}$$

This can be rewritten into the form:

$$\frac{p_p}{p_n} = \frac{n_n}{n_p} = \exp\left(\frac{qV_0}{k_b T}\right) \tag{9.21}$$

Using the expressions in Eq. (9.21), we can write the built-in potential as a function of the doping concentrations:

$$V_0 = \frac{k_b T}{q} \ln\left(\frac{N_A N_D}{n_i^2}\right) \tag{9.22}$$

This potential exists at equilibrium and is a direct consequence of the junction between dissimilarly doped materials. However, it cannot be directly measured using a voltmeter because voltmeters measure the chemical potential difference, and the chemical potential is the same throughout the device since it is at thermal equilibrium with balanced drift and diffusion currents everywhere.

9.2.5 Depletion Width

It is now possible to relate the width W_0 of the space charge region, as well as its extent on either side of the *p-n* junction, with the built-in potential. From the expression of the built-in potential in (Eq. 9.22), we can express the depletion width as:

$$W_0 = \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A + N_D}{N_A N_D}\right) V_0} \tag{9.23}$$

which becomes, after considering Eq. (9.22):

$$W_0 = \sqrt{\frac{2\varepsilon k_b T}{q^2}\left(\frac{N_A + N_D}{N_A N_D}\right) \ln\left(\frac{N_A N_D}{n_i^2}\right)} \tag{9.24}$$

The extent of the depletion width into each side of the *p-n* junction can then be determined by replacing W_0 from Eq. (9.23) into Eq. (9.7):

$$\begin{cases} x_{p0} = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_D}{N_A(N_A+N_D)}\right)V_0} \\ x_{n0} = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_A}{N_D(N_A+N_D)}\right)V_0} \end{cases} \tag{9.25}$$

These last two expressions show that the space charge region extends more into the region of lower doping, in accordance with Subsect. 9.2.2.

Example

Q Consider a GaAs abrupt *p-n* junction with a doping level on the *p*-type side of $N_A = 2 \times 10^{17}$ cm^{-3} and a doping level on the *n*-type side of $N_D = 1 \times 10^{17}$ cm^{-3}. Estimate the depletion region widths on the *p*-type side and the *n*-type side at 300 K.

A The depletion region widths sought are given by the following expressions:

$$\begin{cases} x_{p0} = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_D}{N_A(N_A+N_D)}\right)V_0} \\ x_{n0} = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_A}{N_D(N_A+N_D)}\right)V_0} \end{cases}$$

where ε is the dielectric constant of GaAs ($\varepsilon = 13.1\varepsilon_0$) and V_0 is the built-in potential. The latter is calculated from:

$$\begin{aligned} V_0 &= \frac{k_b T}{q} \ln\left(\frac{N_A N_D}{n_i^2}\right) \\ &= \frac{\left(1.38066 \times 10^{-23}\right) \times 300}{1.60218 \times 10^{-19}} \ln\left(\frac{\left(2 \times 10^{17}\right)\left(1 \times 10^{17}\right)}{\left(1.79 \times 10^{6}\right)^2}\right) \\ &= 1.297\ V \end{aligned}$$

because the intrinsic carrier concentration in GaAs at 300 K is $n_i = 1.79 \times 10^6$ cm^{-3}. The widths can then be calculated as:

$$\begin{aligned} x_{p0} &= \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_D}{N_A(N_A+N_D)}\right)V_0} \\ &= \sqrt{\begin{array}{l}\dfrac{2 \times \left(13.1 \times 8.85418 \times 10^{-14}\right)}{1.60218 \times 10^{-19}} \\ \times\left(\dfrac{1 \times 10^{17}}{\left(2 \times 10^{17}\right)\left(2 \times 10^{17} + 1 \times 10^{17}\right)}\right) \times 1.297\end{array}} \\ x_{p0} &= 5.6 \times 10^{-6}\text{cm} \\ x_{p0} &= 56\ \text{nm} \end{aligned}$$

and

$$\begin{aligned} x_{n0} &= \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A}{N_D(N_A+N_D)}\right)V_0} \\ &= \frac{N_A}{N_D}x_{p0} \\ &= \frac{2\times 10^{17}}{1\times 10^{17}}5.6\times 10^{-6} \\ &= 11.2\times 10^{-6}\text{cm} \\ x_{n0} &= 112\text{ nm} \end{aligned}$$

9.2.6 Energy Band Profile and Fermi Energy

Because of the presence of a built-in potential, the allowed energy bands in the semiconductor, e.g., the conduction and the valence bands in particular, are shifted too. The resulting energy band profile is obtained by multiplying the potential by the charge of an electron ($-q$). This is shown in Fig. 9.10e, where it is conventional to plot the bottom of the conduction band (E_C) and the top of the valence band (E_V) across the *p-n* structure.

The reason why we must multiply by the negative charge of an electron is because the resulting band diagram corresponds to the allowed energy states for electrons. This is intuitively understandable because the electrons are more likely to be where there is a higher positive electrical potential; thus the energy band for electrons will be lower there.

We therefore see that the conduction and valence bands are "bent" from the *p*-type to the *n*-type regions. Moreover, the amount of band bending is directly related to the built-in potential:

$$E_{Vp} - E_{Vn} = E_{Cp} - E_{Cn} = qV_0 \tag{9.26}$$

Example

Q Estimate the energy band bending from the *p*-type side to the *n*-type side in a GaAs abrupt *p-n* junction with a doping level on the *p*-type side of $N_A = 2\times 10^{17}$ cm^{-3} and a doping level on the *n*-type side of $N_D = 1\times 10^{17}$ cm^{-3} at 300 K.

A From the previous example, we know that the built-in potential is $V_0 = 1.297$ V. The band bending is therefore equal to $qV_0 = 1.297$ eV.

Away from the space charge region, the Fermi energies in the *p*-type and *n*-type regions are denoted E_{Fp} and E_{Fn}, respectively, as shown in Fig. 9.8 . At equilibrium, these quantities must be equal. Indeed, the hole density in the *p*-type and *n*-type regions is given by Eq. (7.29) in the nondegenerate case:

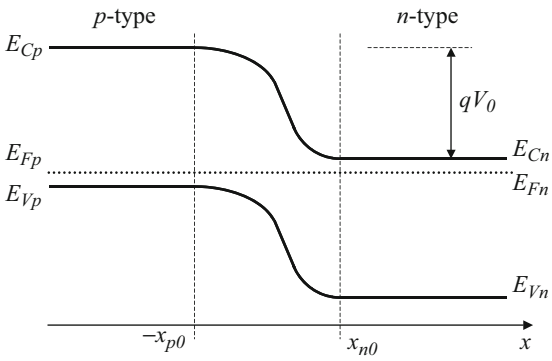


Fig. 9.8 Energy band profile across a *p-n* junction. This profile is obtained by multiplying the potential in Fig. 9.6 by −q, the electrical charge of electrons

$$\begin{cases} p_{\mathrm{p}} = N_{\mathrm{v}} \exp\left(\dfrac{E_{\mathrm{Vp}} - E_{\mathrm{Fp}}}{k_{\mathrm{b}}T}\right) \\ p_{\mathrm{n}} = N_{\mathrm{v}} \exp\left(\dfrac{E_{\mathrm{Vn}} - E_{\mathrm{Fn}}}{k_{\mathrm{b}}T}\right) \end{cases} \tag{9.27}$$

Utilizing Eq. (9.25), we get:

$$\exp\left(\frac{qV_0}{k_{\mathrm{b}}T}\right) = \frac{p_{\mathrm{p}}}{p_{\mathrm{n}}} = \frac{\exp\left(\frac{E_{\mathrm{Vp}} - E_{\mathrm{Fp}}}{k_{\mathrm{b}}T}\right)}{\exp\left(\frac{E_{\mathrm{Vn}} - E_{\mathrm{Fn}}}{k_{\mathrm{b}}T}\right)}$$

$$\exp\left(\frac{qV_0}{k_{\mathrm{b}}T}\right) = \exp\left(\frac{E_{\mathrm{Vp}} - E_{\mathrm{Vn}}}{k_{\mathrm{b}}T}\right) \exp\left(\frac{E_{\mathrm{Fn}} - E_{\mathrm{Fp}}}{k_{\mathrm{b}}T}\right) \tag{9.28}$$

In addition, by using Eq. (9.26) in this expression, we get:

$$1 = \exp\left(\frac{E_{\mathrm{Fn}} - E_{\mathrm{Fp}}}{k_{\mathrm{b}}T}\right)$$

which means that $E_{\mathrm{Fn}} = E_{\mathrm{Fp}}$, i.e., the Fermi energies in the *p*-type and *n*-type regions are equal, and this has already been anticipated in Fig. 9.8 . In fact, this is a general and important property that, at thermal equilibrium, the Fermi energies of dissimilar materials must be equal. This physically means that there must not be a net flow of holes or electrons across the structure at equilibrium.

9.3 Non-equilibrium Properties of *p-n* Junctions

The most interesting and practical properties of a *p-n* junction are observed under non-equilibrium conditions, such as when a voltage is applied across it and/or when it is illuminated. Because of its nonsymmetrical nature, a *p-n* junction will exhibit different properties depending on the polarity of the external voltage or bias applied.

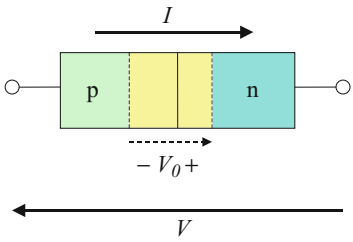


Fig. 9.9 Convention for the polarity of the external voltage and current

The sign convention used for the external voltage and the current in a *p-n* junction is shown in Fig. 9.9: the voltage will be positive if the applied potential on the *p*-type side is higher than that applied on the *n*-type. Note that the built-in voltage V_0 has been taken to be positive.

When an external bias is applied, the diffusion and drift currents do not balance each other anymore. This imbalance results in a net flow of electrical current in one or the other direction. In addition, the internal electric field and voltage across the *p-n* junction, the depletion width, and the energy band profile will all be changed. In this section, we will review how these parameters are modified.

9.3.1 Forward Bias: A Qualitative Description

When an external bias V is applied to the *p-n* structure depicted in Fig., there is usually some voltage drop across both the neutral bulk *p*-type and the *n*-type regions (i.e., outside the space charge region) due to Ohm's law (Sect. 8.2). In other words, the entire external bias is not applied across the transition region because part of it would be "lost" across the neutral regions due to their electrical resistance.

However, in most semiconductor devices which use *p-n* junctions, the length of these neutral regions which the electrical current would have to flow through is small, and any voltage drop would thus be negligible compared to the voltage change across the transition region. In our discussion, for now we will assume that the external bias is applied directly to the limits of the space charge region.

According to the sign convention in Fig. 9.10d, the total voltage across the transition region is now given by V_0-V. There are typically two regimes which need to be considered for the non-equilibrium conditions of a *p-n* junction: forward bias and reverse bias.

In the forward bias regime, corresponding to $V > 0$, the total voltage or potential barrier across the transition region is actually reduced from V_0 to $V_0 - V$, which has a number of consequences. First, the strength of the internal electric field associated with the lower potential barrier is reduced as well, as shown in Fig. 9.10c. This in turn means that the width of the space charge region is reduced because fewer electrical charges are needed to maintain this electric field, as shown in Fig. 9.10b. In other words, W_0 is reduced and is now denoted W, x_{n0} becomes x_n, and x_{p0} becomes x_p, as illustrated in Fig. 9.10a. As the internal voltage is reduced from its

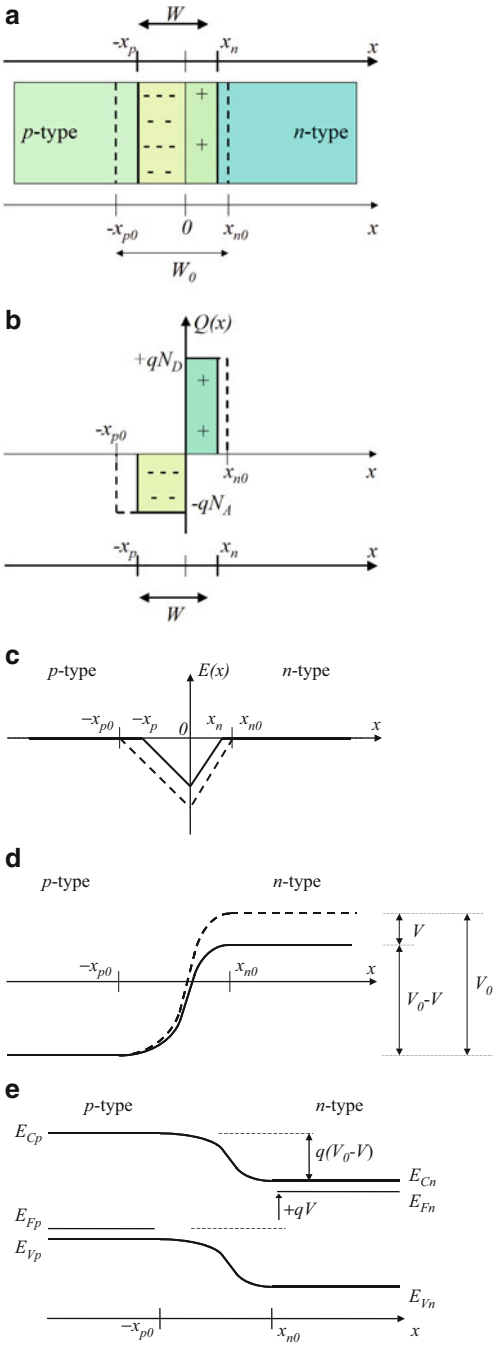


Fig. 9.10 (**a**) Space charge region width, (**b**) electrical charge density, (**c**) electric field strength, (**d**) potential profile, and (**e**) energy band profile of a *p-n* junction under forward bias ($V > 0$). The thick dashed curves represent the equilibrium case for comparison

equilibrium value by an amount equal to V, the energy band profile is changed, and the amount of band bending is reduced by qV, as depicted in Fig. 9.10e. This means that:

$$E_{\mathrm{Vp}} - E_{\mathrm{Vn}} = E_{\mathrm{Cp}} - E_{\mathrm{Cn}} = q(V_0 - V) \tag{9.29}$$

instead of Eq. (9.26). Furthermore, we can still consider that the Fermi energy levels outside the space charge region, i.e., in the neutral bulk p-type (E_{Fp}) and n-type (E_{Fn}) regions, are located at their equilibrium positions because we assumed no voltage drop in these regions. Therefore, because the band bending has been reduced by qV, according to Fig. (e), we must have:

$$E_{\mathrm{Fp}} - E_{\mathrm{Fn}} = -qV \tag{9.30}$$

This means that the Fermi energy is not constant throughout the p-n junction structure, but the Fermi energy levels in the neutral p-type and the n-type regions are separated by qV, where V is the applied external bias. This is a direct consequence of a non-equilibrium condition.

Let us now qualitatively examine the effects of a forward bias on the diffusion and drift currents across the space charge region of a p-n junction. As we saw in the previous section, the diffusion current arises from the difference between the densities of charge carriers on either side of the junction area. It corresponds to the motion of electrons from the n-type region toward the p-type region, and conversely for holes. This means that, at its origin, the diffusion current is related to the motion of majority carriers (e.g., electrons in the n-type region). However, as soon as these carriers reach the other side of the junction, they become minority carriers. Therefore, the diffusion current acts as if it injects minority carriers into one side of the junction by pulling them from the other side of the junction where they are majority carriers.

At equilibrium, the diffusion process is stabilized when the built-in electric field exerts a force that exactly counterbalances the diffusion of these charge carriers. Under a forward bias, as we just saw in Fig. 9.10c, this electric field strength is reduced. Therefore, each type of charge carriers can diffuse more easily, which means that the diffusion currents for both types of carrier increase under a forward bias.

This can also be understood by examining the energy band profile. For example, when the electrons in the n-type region, on the right-hand side of Fig. 9.10c where they are more concentrated, diffuse toward the p-type region where they are less concentrated, the allowed energy states are located at higher energies. This means that the diffusion electrons have to cross a high-energy barrier. Under a forward bias, this energy barrier is reduced, as shown in Fig. 9.10e, and more electrons can thus participate in the diffusion toward the p-type region. A similar argument is valid for holes. As a result, *the diffusion currents for both types of carrier increase under a forward bias*.

By contrast, the *drift current does not change with an external bias*, although this may seem contradictory with the fact that the internal electric field is weaker. This can be understood by examining the drift current in more detail. We saw in Sect. 9.2

that the drift current counterbalanced the diffusion of charge carriers and thus consisted of electrons moving toward the *n*-type region and holes moving toward the *p*-type region. This means that, at its origin, the drift current is related to the motion of minority carriers, such as electrons in the *p*-type region which drift toward the *n*-type region under the influence of the electric field. The drift current thus plays the converse role of the diffusion current. The drift current acts as if it extracts minority carriers from one side of the junction to send them to the other side of the junction where they are majority carriers. Because the concentrations of minority carriers are very small (see Eq. (9.2)), the drift currents are mostly limited by the number of minority carriers available for drift (i.e., electrons on the *p*-type region and holes on the *n*-type region) rather than by the speed at which they would drift (i.e., the strength of the electric field). We then understand why the drift current does not change significantly when an external bias is applied, in comparison to the diffusion current.

9.3.2 Reverse Bias: A Qualitative Description

By contrast, in the reverse bias regime, corresponding to $V < 0$, the total voltage or potential barrier across the transition region is actually increased from V_0 to $V_0 - V$, which also has the opposite effects of a forward bias. The strength of the internal electric field is increased, as shown in Fig. (c). This enlarges the width of the space charge region from W_0 to W (with x_{n0} becoming x_n, and x_{p0} becoming x_p, as illustrated in Fig. 9.11a) because more electrical charges are needed to maintain this electric field, as shown in Fig. 9.11b. As the internal voltage is increased from its equilibrium value by an amount equal to $-V$, the energy band profile is changed, and the amount of band bending is increased by $-qV$, as depicted in Fig. 9.11e. The total amount of band bending is still given by the expression in Eq. (9.29). The difference between the Fermi energy levels outside the space charge region is also still given by Eq. (9.30).

In addition, by contrast with the forward bias case, *the diffusion currents for both types of carrier decrease under a reverse bias*. However *the drift current still does not change significantly* in comparison to the diffusion current when a reverse bias is applied, for the same reason as discussed previously.

9.3.3 A Quantitative Description

In the previous subsections, we have expressed quantitatively the amount of band bending and the difference between the Fermi energy levels of the neutral *p*-type and *n*-type regions as a function of the applied external bias (Eqs. (9.29) and (9.30), respectively).

In fact, most of the relations that were derived in Sect. 9.2 for the equilibrium case are valid when an external bias voltage V is applied, provided we make the following transformations:

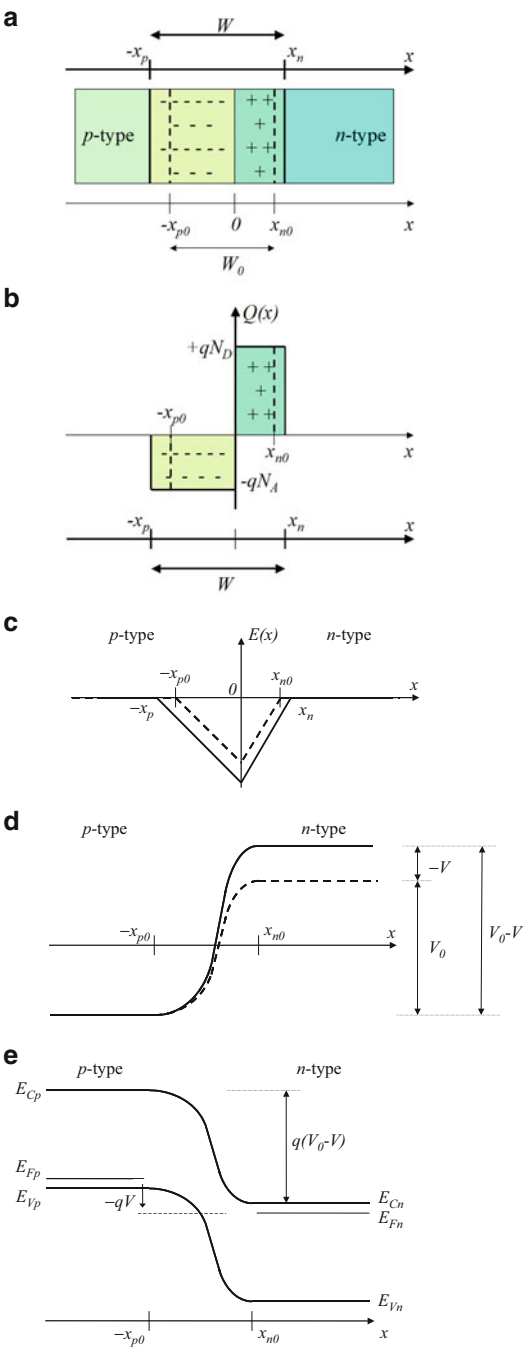


Fig. 9.11 (**a**) Space charge region width, (**b**) electrical charge density, (**c**) electric field strength, (**d**) potential profile, and (**e**) energy band profile of a *p-n* junction under reverse bias ($V < 0$). The thick dashed curves represent the equilibrium case for comparison

$$\begin{cases} W_0 & \rightarrow & W \\ x_{p0} & \rightarrow & x_p \\ x_{n0} & \rightarrow & x_n \\ V_0 & \rightarrow & V_0 - V \end{cases} \tag{9.31}$$

This statement is justified by the fact that most of the expressions in Sect. 9.2 have been obtained without invoking the equilibrium condition of Eq. (9.3) but by using the electrical charge neutrality principle and Gauss's law instead which are valid at all times.

The following few relations will be important for future discussions. The depletion width can be obtained from Eq. (9.23) by using Eq. (9.31):

$$W = \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)(V_0 - V)} \tag{9.32}$$

for $V < V_0$. We clearly see that the depletion width shrinks when a forward bias is applied ($V > 0$), whereas it expands when a reverse bias is applied ($V < 0$). This confirms the qualitative discussion of the previous subsection.

Example

Q Calculate the ratio of the depletion region width W under a forward bias of 0.3 V to the equilibrium width W_0, for a GaAs abrupt *p-n* junction with a doping level on the *p*-type side of $N_A = 2 \times 10^{17}$ cm^{-3} and a doping level on the *n*-type side of $N_D = 1 \times 10^{17}$ cm^{-3} at 300 K.

A The depletion width W under a bias V is given by the expression: $W = \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A+N_D}{N_A N_D}\right)(V_0 - V)}$, where the built-in potential is $V_0 = 1.297$ V, as determined in earlier examples. The ratio sought is therefore:

$$\frac{W}{W_0} = \sqrt{\frac{(V_0 - V)}{V_0}} = \sqrt{\frac{(2.297 - 0.3)}{1.297}} = 0.877$$

The depletion width is then:

$$\begin{aligned} W &= 0.877 W_0 = 0.877\left(x_{p0} + x_{n0}\right) \\ &= 0.877(56 + 112) \\ &= 147 \text{ nm} \end{aligned}$$

The extent of the space charge region inside the *p*-type and *n*-type regions, as shown in Figs. 9.9a and 9.10a, can be obtained from Eq. (9.25):

$$\begin{cases} x_p = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_D}{N_A(N_A+N_D)}\right)(V_0 - V)} \\ x_n = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_A}{N_D(N_A+N_D)}\right)(V_0 - V)} \end{cases} \tag{9.33}$$

Similarly, the non-equilibrium hole and electron concentrations at the edges of the space charge region, denoted $p(-x_p)$, $p(x_n)$, $n(-x_p)$, and $n(x_n)$, can be obtained by considering Eq. (9.21):

$$\frac{p(-x_p)}{p(x_n)} = \frac{n(x_n)}{n(-x_p)} = \exp\left(\frac{q(V_0 - V)}{k_b T}\right) \tag{9.34}$$

In addition, following our previous discussion, we realize that the majority carrier concentrations are little changed under a moderate forward or a reverse bias, i.e., $p(-x_p) = p_p$ and $n(-x_n) = n_n$, which after replacing in Eq. (9.34) to:

$$\frac{p_p}{p(x_n)} = \frac{n_n}{n(-x_p)} = \exp\left(\frac{q(V_0 - V)}{k_b T}\right)$$

and by using Eq. (9.21) to eliminate p_p and n_n from this latest equation:

$$\frac{p(x_n)}{p_n} = \frac{n(-x_p)}{n_p} = \exp\left(\frac{qV}{k_b T}\right) \tag{9.35}$$

These expressions are important as they show that, when an external bias voltage is applied, the *minority carrier concentrations* at the boundary of the space charge region, $p(x_n)$ and $n(x_p)$, are directly and simply related to the equilibrium minority carrier concentrations p_n and n_p, and the applied bias voltage V. All these relations will prove important in the derivation of the diode equation for an ideal *p-n* junction which will be the topic of the next subsection.

Example

Q Calculate the minority carrier concentrations at x_n and $-x_p$ for the GaAs *p-n* junction described in the previous example.

A The minority carrier concentrations at x_n and $-x_p$ are given by: $\frac{p(x_n)}{p_n} = \frac{n(-x_p)}{n_p} = \exp\left(\frac{qV}{k_b T}\right)$, where p_n and n_p are the minority carrier concentrations in the neutral *n*-type side and *p*-type side, respectively, at equilibrium. These are given by the action mass law:

$$p_n = \frac{n_i^2}{N_D} = \frac{\left(1.79 \times 10^6\right)^2}{1 \times 10^{17}} = 3.20 \times 10^{-5}\text{cm}^{-3}\text{and}$$

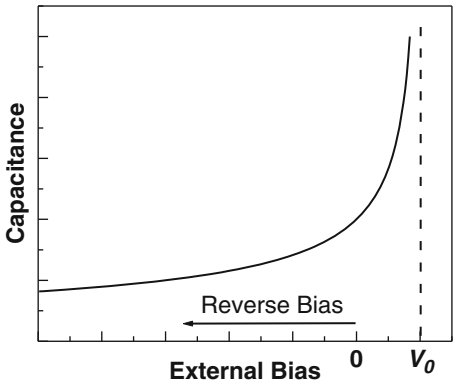


Fig. 9.12 Depletion layer capacitance as a function of bias voltage, showing the increase in capacitance with forward bias and the decrease with reverse bias

$$n_{\mathrm{p}} = \frac{n_{\mathrm{i}}^2}{N_{\mathrm{A}}} = \frac{\left(1.79 \times 10^6\right)^2}{2 \times 10^{17}} = 1.60 \times 10^{-5}\mathrm{cm}^{-3}..$$

In addition, the exponential is numerically equal to:

$\exp\left(\frac{qV}{k_{\mathrm{b}}T}\right) = \exp\left(\frac{\left(1.60218\times10^{-19}\right)\times0.3}{\left(1.38066\times10^{-23}\right)\times300}\right) = 1.1 \times 10^5$. Thus, we get:

$$p(x_{\mathrm{n}}) = \left(3.20 \times 10^{-5}\right)\left(1.1 \times 10^5\right) \approx 3.5\ \mathrm{cm}^{-3} \text{ and}$$

$$n(x_{\mathrm{p}}) = \left(1.60 \times 10^{-5}\right)\left(1.1 \times 10^5\right) \approx 1.76\,\mathrm{cm}^{-3}$$

9.3.4 Depletion Layer Capacitance

The depletion layer is relatively devoid of mobile carriers and can therefore be thought of as somewhat similar to the dielectric in a capacitor. Positive and negative charges are separated by this depletion layer, and this leads to a capacitance associated with the *p-n* junction. This capacitance can be thought of as like that of a parallel plate capacitor and expressed as:

$$C_{\mathrm{dep}} = \frac{\varepsilon A}{W} \tag{9.36}$$

However rather than being constant, the capacitance of a *p-n* junction varies with the reverse bias via the voltage dependence of the depletion width as shown in Fig. 9.12.

More formally, the capacitance of the *p-n* junction can be derived starting from the definition of capacitance:

$$C_{\mathrm{dep}} = \left|\frac{dQ}{dV}\right| \tag{9.37}$$

where dQ is the incremental change in charge stored on either side of the junction for an incremental increase in voltage of dV. For the abrupt junction, the charge stored on either side of the junction can be expressed as:

$$Q_{\text{dep}} = qAN_{\text{D}}x_{\text{n}} = qAN_{\text{A}}x_{\text{p}} \tag{9.38}$$

where x_{n} and x_{p} are given by Eq. (9.33). Substituting in Eq. (9.38) for either term gives the equation:

$$Q_{\text{dep}} = A\sqrt{2q\varepsilon\frac{N_{\text{A}}N_{\text{D}}}{(N_{\text{A}}+N_{\text{D}})}(V_0 - V)}$$

which can then be differentiated with respect to V to yield:

$$C_{\text{dep}} = \frac{A}{\sqrt{(V_0 - V)}}\sqrt{\frac{q\varepsilon}{2}\frac{N_{\text{A}}N_{\text{D}}}{(N_{\text{A}}+N_{\text{D}})}} \tag{9.39}$$

which we can see reduces to Eq. (9.36) above when $V = 0$.

The voltage dependence of the *p-n* junction capacitance is used in varactor diodes or varicaps, in tuning circuits where the diode is reverse-biased to prevent forward conduction, and a small DC tuning voltage is applied to vary the capacitance. Additionally, measuring the capacitance of a diode as a function of bias can be used to extract information about the built-in voltage and the doping profile. This can be done by plotting $1/C_{\text{dep}}$ vs. applied voltage:

$$V = A^2\left[\frac{q\varepsilon(N_{\text{A}}N_{\text{D}})}{2(N_{\text{A}}+N_{\text{D}})}\right]\frac{1}{C_{\text{dep}}^2} - V_o \tag{9.40}$$

In the case of an abrupt one-sided junction (such as a p^+n^- or a metal-semiconductor Schottky diode (see Sect. 9.5)), this equation reduces further, and the carrier concentrations can be extracted more directly:

$$\begin{aligned} V &= \frac{A^2q\varepsilon}{2}N_{\text{A}}\frac{1}{C_{\text{dep}}^2} - V_o, \quad (N_{\text{D}} >> N_{\text{A}}) \\ V &= \frac{A^2q\varepsilon}{2}N_{\text{D}}\frac{1}{C_{\text{dep}}^2} - V_o, \quad (N_{\text{A}} >> N_{\text{D}}) \end{aligned} \tag{9.41}$$

9.3.5 Ideal p-n Junction Diode Equation

The diode equation refers to the mathematical expression which relates the total electrical current I through an ideal *p-n* junction to the applied external bias voltage V. It is also referred as the current-voltage or *I-V* characteristic of the diode. To determine it, we must focus our analysis on the minority carriers, i.e., holes in the *n*-type region and electrons in the *p*-type region.

In addition to the depletion approximation model considered so far, a few more assumptions need to be considered:

(i) First, we assume that there are no external sources of carrier generation.
(ii) No recombination of charge carriers occurs within the space charge region.
(iii) We assume that the applied biases are moderate enough to ensure that the minority carriers remain much less numerous than the majority carriers in the neutral regions.
(iv) Finally, we assume that the change in minority carrier concentrations in the neutral regions does not result in a non-negligible electric field.

In virtue of assumptions (i) and (ii), any hole or electron that has diffused across the space charge region must be present at its boundaries, i.e., at $-x_p$ and x_n, respectively. When a bias V is applied, the concentrations of these holes and electrons, which are in excess of their equilibrium concentrations, are given by:

$$\begin{cases} \Delta p_n = p(x_n) - p_n \\ \Delta n_p = n(-x_p) - n_p \end{cases}$$

This becomes after using Eq. (9.35):

$$\begin{cases} \Delta p_n = p_n\left(e^{\frac{qV}{k_b T}} - 1\right) \\ \Delta n_p = n_p\left(e^{\frac{qV}{k_b T}} - 1\right) \end{cases} \tag{9.42}$$

Here, and in the rest of the text, we will use the extended meaning of the term "excess carrier." For example, if Δp_n and Δn_p are positive, i.e., $V > 0$ or forward bias, then there are net real excesses of holes and electrons at the space charge boundaries, and we talk about minority carrier injection. This is shown in Fig. 9.13.

But if Δp_n and Δn_p are negative, i.e., $V < 0$ or reverse bias, then there are net real deficiencies of holes and electrons, and we talk about minority carrier extraction. In this case, the minority carriers at the boundaries of the space charge region are less numerous than in the bulk neutral material; therefore there is a diffusion of minority carriers from the bulk neutral region toward the edges of the space charge region. This is illustrated in Fig. 9.14.

Returning to the forward bias case, the excess holes, present at $x = x_n$ with a concentration Δp_n, will be diffusing deeper into the neutral n-type region where their equilibrium concentration is only p_n. As they diffuse, they will experience recombination as discussed in Chap. 8, with a characteristic diffusion length L_p in the steady-state regime. The excess hole concentration is therefore reduced as we advance deeper in the material. This situation has already been encountered in Chap. 8 and the analytical expression for $\delta p_n(x_1)$, the excess hole concentration at a position x_1, is obtained for Eq. (8.55):

$$\delta p_n(x_1) = \Delta p_n e^{-\frac{x_1}{L_p}} \tag{9.43}$$

where L_p is the hole diffusion length in the n-type region. In this expression, we chose another axis, denoted x_1, oriented in the same direction as the original axis x and with its origin at $x = x_n$. It is important to remember that the excess

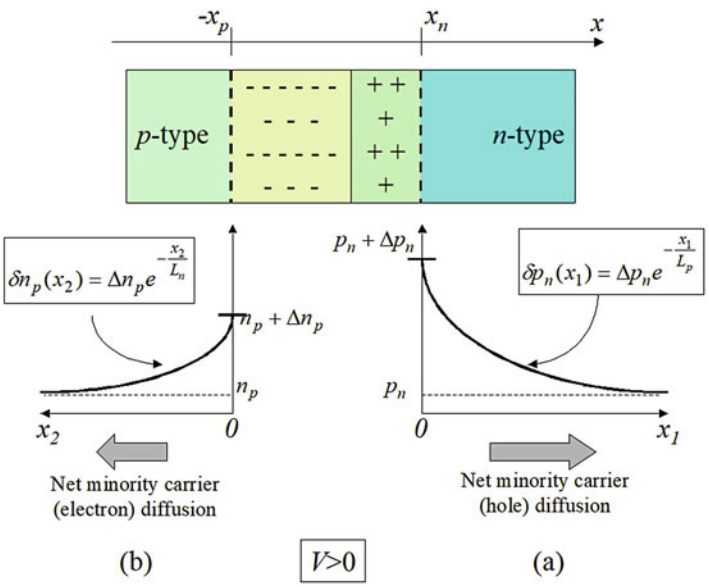


Fig. 9.13 (**a**) Excess hole concentration profile in the *n*-type region, and (**b**) excess electron concentration profile in the *p*-type region, under a forward bias. The excess carrier concentrations decrease, following an exponential decay, as they go further from the edges of the depletion region

concentration of holes at $x = x_n$ remains constant at Δp_n given by Eq. (9.42) because holes are continuously injected or extracted through the space charge region into or from the *n*-type region due to the application of the external bias voltage. We can make use of Fig. 8.7 to plot the spatial profile of the excess hole concentration in Fig. 9.13a for the forward bias case and Fig. 9.14a for the reverse bias case.

Conversely, the excess electrons present at $x = -x_p$ with a concentration Δn_p will diffuse deeper into the neutral *p*-type region, with a diffusion length L_n. This leads to the spatial profile $\delta n_p(x_2)$ shown in Fig. 9.13b for the forward bias case and Fig. 9.14b for the reverse bias case, and it is analytically given by:

$$\delta n_p(x_2) = \Delta n_p e^{-\frac{x_2}{L_n}} \tag{9.44}$$

where L_n is the electron diffusion length in the *p*-type region. It is important to note that, here, we chose the sign convention for the axis x_2 in the opposite direction of the original axis x because the electrons diffuse in this opposite direction.

There are essentially two methods to compute the diode equation. The first one consists of analyzing the diffusion currents in the *p-n* junction. From our discussion in Subsect. 9.3.1, we understand that, when an external bias is applied, the drift currents across the space charge region do not vary, whereas the diffusion currents change. The sum of the increments in the hole and the electron diffusion currents across the space charge region is thus a direct measure of the net electrical current through the *p-n* junction since no net current is originally present at equilibrium,

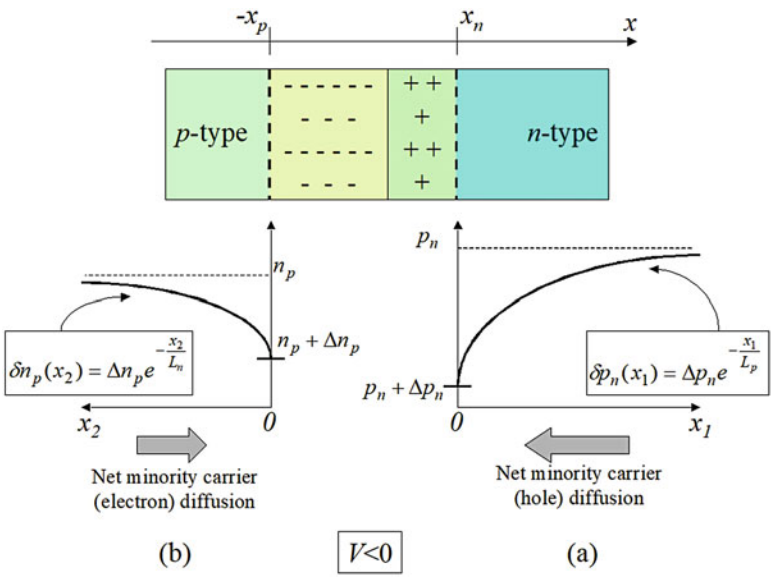


Fig. 9.14 (**a**) "Excess" hole concentration profile in the *n*-type region, and (**b**) "excess" electron concentration profile in the *p*-type region, under a reverse bias. These carrier concentrations change following an exponential dependence as they go further away from the edges of depletion region

because we have assumed there are no external sources of carrier generation and because the total electrical current is constant throughout a two-terminal device, such as the *p-n* junction earlier shown in Fig. 9.8.

The incremental diffusion currents are the diffusion currents which result from the excess carriers in the material. The diffusion current densities for electrons and holes can be obtained from Eqs. (8.36) and (8.38) and are given by:

$$\begin{cases} J_{\mathrm{h}}^{\mathrm{diff}}(x_1) = -qD_{\mathrm{p}}\dfrac{d(\delta p_{\mathrm{n}}(x_1))}{dx_1} \\ J_{\mathrm{e}}^{\mathrm{diff}}(x_2) = qD_{\mathrm{n}}\dfrac{d\left(\delta n_{\mathrm{p}}(x_2)\right)}{dx_2} \end{cases} \tag{9.45}$$

Using the expressions of the excess carrier concentrations in Eqs. (9.43) and (9.44), we get:

$$\begin{cases} J_{\mathrm{h}}^{\mathrm{diff}}(x_1) = +q\dfrac{D_{\mathrm{p}}}{L_{\mathrm{p}}}\Delta p_{\mathrm{n}} e^{-\frac{x_1}{L_p}} \\ J_{\mathrm{e}}^{\mathrm{diff}}(x_2) = -q\dfrac{D_{\mathrm{n}}}{L_{\mathrm{n}}}\Delta n_{\mathrm{p}} e^{-\frac{x_2}{L_n}} \end{cases} \tag{9.46}$$

In order to obtain the total current through the *p-n* junction, we must evaluate the diffusion current densities for holes and electrons at the limits of the space charge region at $x = x_n$ and $x = -x_p$, respectively, or equivalently at $x_1 = x_2 = 0$:

$$\begin{cases} J_h^{diff}(0) = +q\dfrac{D_p}{L_p}\Delta p_n \\ J_e^{diff}(0) = -q\dfrac{D_n}{L_n}\Delta n_p \end{cases} \tag{9.47}$$

Example

Q Estimate the ratio of the diffusion current densities of holes and electrons for the GaAs *p-n* junction described in the previous example.

A The ratio of the diffusion currents is given by: $\left|\dfrac{J_h^{diff}(0)}{J_e^{diff}(0)}\right| = \dfrac{D_p}{D_n}\dfrac{L_n}{L_p}\dfrac{\Delta p_n}{\Delta n_p}$, where Δp_n and Δn_p are the excess minority carrier concentrations at the limits of the depletion region. These quantities are given by: $\Delta p_n = p_n\left(e^{\frac{qV}{k_bT}} - 1\right)$ and $\Delta n_p = n_p\left(e^{\frac{qV}{k_bT}} - 1\right)$. Their ratio is then: $\dfrac{\Delta p_n}{\Delta n_p} = \dfrac{p_n}{n_p} = \dfrac{{}^{n_i^2}/_{N_D}}{{}^{n_i^2}/_{N_A}} = \dfrac{N_A}{N_D}$. In addition, the diffusion lengths can be expressed as a function of the minority carrier lifetime on the *n*-type and the *p*-type sides. These lead to the ratio: $\left|\dfrac{J_h^{diff}(0)}{J_e^{diff}(0)}\right| = \dfrac{D_p}{D_n}\dfrac{\sqrt{D_n\tau_n}}{\sqrt{D_p\tau_p}}\dfrac{N_A}{N_D}$. Assuming that the minority carrier lifetimes are the same for holes and electrons, we get: $\left|\dfrac{J_h^{diff}(0)}{J_e^{diff}(0)}\right| = \sqrt{\dfrac{D_p}{D_n}}\dfrac{N_A}{N_D}$. The ratio of the diffusion coefficients can be calculated using the majority carrier mobilities through the Einstein relations and we obtain: $\left|\dfrac{J_h^{diff}(0)}{J_e^{diff}(0)}\right| = \sqrt{\dfrac{\mu_h}{\mu_e}}\dfrac{N_A}{N_D}$ and

$$\left|\frac{J_h^{diff}(0)}{J_e^{diff}(0)}\right| = \sqrt{\frac{400}{8500}}\frac{2\times10^{17}}{1\times10^{17}} \approx 0.43$$

In all these expressions of current densities, it is important to remember that the sign convention for the current density $J_h^{diff}(x_1)$ is the same as the axis x, whereas for $J_e^{diff}(x_2)$ it is opposite that of axis x. The total current density is the sum of the hole and electron diffusion currents, with however a sign difference:

$$J_{total} = J_h^{diff}(0) - J_e^{diff}(0) \tag{9.48}$$

The minus sign for $J_e^{diff}(0)$ accounts for the sign convention chosen for axis x_2. Inserting Eq. (9.47) into this relation, we get:

$$J_{\text{total}} = q\left(\frac{D_\text{p}}{L_\text{p}}\Delta p_\text{n} + \frac{D_\text{n}}{L_\text{n}}\Delta n_\text{p}\right) \tag{9.49}$$

and using Eq. (9.42), we finally obtain:

$$J_{\text{total}} = q\left(\frac{D_\text{p}}{L_\text{p}}p_\text{n} + \frac{D_\text{n}}{L_\text{n}}n_\text{p}\right)\left(e^{\frac{qV}{k_\text{b}T}} - 1\right) \tag{9.50}$$

The total current is given by the total current density multiplied by the area of the *p-n* junction. If we assume a uniform area A, we get:

$$I_{\text{total}} = AJ_{\text{total}} = qA\left(\frac{D_\text{p}}{L_\text{p}}p_\text{n} + \frac{D_\text{n}}{L_\text{n}}n_\text{p}\right)\left(e^{\frac{qV}{k_\text{b}T}} - 1\right) \tag{9.51}$$

By introducing a new term I_0, this can be rewritten as:

$$I_{\text{total}} = I_0\left(e^{\frac{qV}{k_\text{b}T}} - 1\right) \tag{9.52}$$

with:

$$I_0 = qA\left(\frac{D_\text{p}}{L_\text{p}}p_\text{n} + \frac{D_\text{n}}{L_\text{n}}n_\text{p}\right) \tag{9.53}$$

Equations (9.52) and (9.53) represent the diode equation for an ideal *p-n* junction. This function is plotted in Fig. 9.15.

We see that under a forward bias, the current increases exponentially as a function of applied voltage. By contrast, under reverse bias, the current rapidly tends toward $-I_0$. The value of the current I_0 is therefore called the reverse saturation current. The physical meaning of this current can be understood as follows. When a strong reverse bias is applied ($V < 0$), the density of minority carriers at the boundary of the space charge region quickly falls to zero according to Eq. (9.35). This means that, inside the depletion region, there is no diffusion of carriers, but only drift currents are

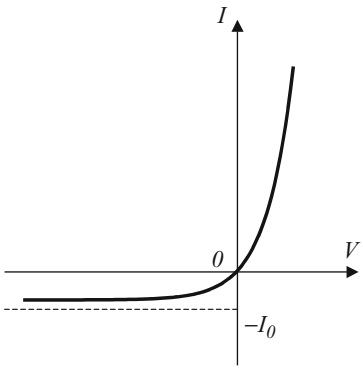


Fig. 9.15 Current-voltage characteristic for an ideal *p-n* junction diode. The dependence of the current on the voltage follows an exponential expression. The current is zero when the voltage is zero, without external excitation

present. Outside the depletion region however, the only charge motion is the diffusion of minority carriers from the neutral regions toward the depletion region, as illustrated by the block arrows in Fig. 9.14. We can therefore say that the saturation current in Eq. (9.53) corresponds to the total drift, across the space charge region, of minority carriers which have been extracted or able to reach the limits of the space charge region through diffusion from the neutral regions.

The *p-n* junction diode acts like a one-way device: when it is forward-biased, current can flow from the *p*-type to the *n*-type region without much resistance, whereas when it is reverse-biased, a very large resistance prevents the current from flowing in the opposite direction from the *n*-type to the *p*-type region.

The second method which can be used to determine the diode equation consists of calculating the total charge accumulated on each side of the junction area. This second method is called the charge control approximation. Let Q_p be the steady-state excess positive charge in the *n*-type region which is given by integrating Eq. (9.43):

$$Q_p = qA \int_0^\infty \delta p_n(x_1) dx_1 = qA\Delta p_n \int_0^\infty e^{-\frac{x_1}{L_p}} dx_1$$

i.e.:

$$Q_p = qAL_p \Delta p_n \tag{9.54}$$

where A is the area of the *p-n* junction. This excess charge is illustrated in Fig. 9.16a, in the forward bias case. The hole diffusion current must then be able to maintain this excess positive charge, even though the holes are recombining. As the average lifetime of holes in the *n*-type region is the recombination lifetime τ_p defined in Subsect. 8.5.3, the hole diffusion current must be able to supply Q_p positive charges during a time equal to τ_p. This current must therefore be $I_p = \frac{Q_p}{\tau_p}$.

Similarly, the excess negative charge in the *p*-type region is given by:

$$Q_n = qAL_n \Delta n_p \tag{9.55}$$

and is shown in Fig. 9.16b. The electron diffusion current into the *p*-type region is $-I_n = -\frac{Q_n}{\tau_n}$. In this last expression, we made use of the same sign convention as for axis x_2. The total current is therefore given by:

$$I_{total} = I_p + I_n = qA\frac{L_p}{\tau_p}\Delta p_n + qA\frac{L_n}{\tau_n}\Delta n_p$$

or:

$$I_{total} = qA\left(\frac{L_p}{\tau_p}\Delta p_n + \frac{L_n}{\tau_n}\Delta n_p\right) \tag{9.56}$$

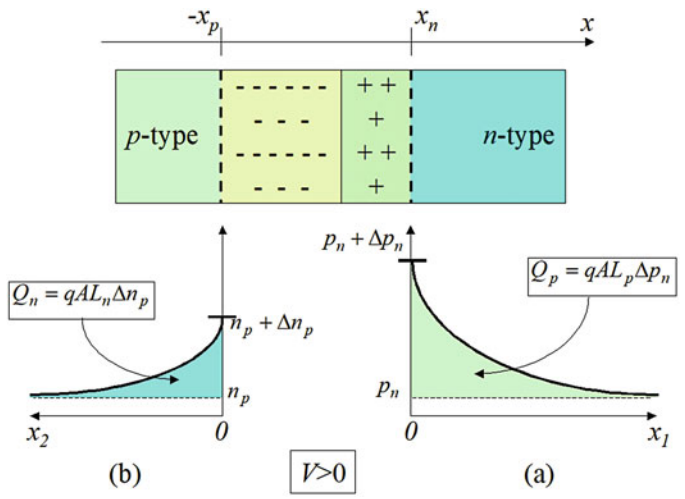


Fig. 9.16 (**a**) Excess positive charge in the *n*-type region and (**b**) excess negative charge in the *p*-type region, under a forward bias. The total excess charges are calculated by integrating the excess carrier concentrations over the volume of the regions outside the depletion region

Using the definition of the diffusion lengths given in Eqs. (8.53) and (8.56), and using Eq. (9.42), we can transform this last expression into:

$$I_{\text{total}} = AJ_{\text{total}} = qA\left(\frac{D_{\text{p}}}{L_{\text{p}}}p_{\text{n}} + \frac{D_{\text{n}}}{L_{\text{n}}}n_{\text{p}}\right)\left(e^{\frac{qV}{k_{\text{b}}T}} - 1\right)$$

and thus get the diode equation obtained in Eq. (9.51).

9.3.6 Minority and Majority Carrier Currents in Neutral Regions

In the previous discussion, we saw that the total electrical current through a *p-n* junction device was determined by the diffusion currents across the space charge region which result in minority carriers being injected into or extracted from the neutral regions under the influence of an applied external bias.

For the sake of clarity, let us consider the example of a forward-biased *p-n* junction, as the one shown in Fig. 9.13. We saw that the excess minority carriers diffuse into the neutral regions following an exponential decay given in Eqs. (9.43) and (9.44). This leads to diffusion currents which also follow an exponential decay, as obtained in Eq. (9.46). However, we know that the total electrical current throughout a two-terminal device is constant. Therefore, the decrease in diffusion current, for example, that of holes in the right-hand side of the figure, as we move away from the space charge region has to be compensated by another current. This is achieved through the drift of majority carriers, for example, electrons in the neutral *n*-type region. Indeed, through their diffusion and recombination, the minority

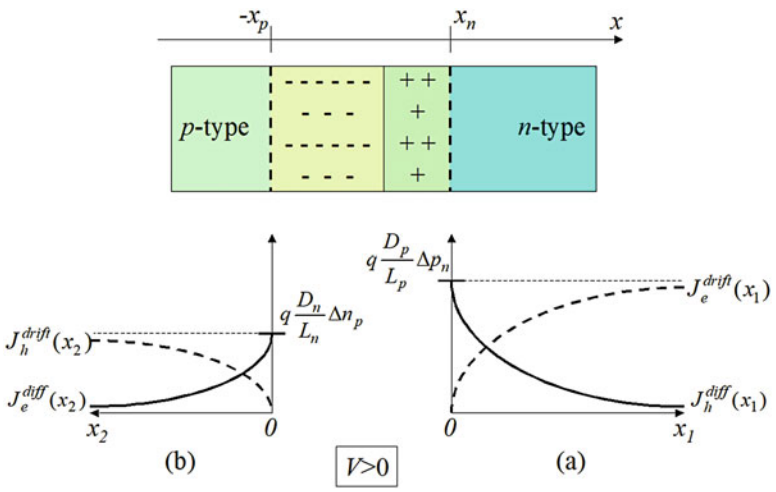


Fig. 9.17 Diffusion current of minority carriers and drift current of majority carriers in the (**a**) *n*-type region and (**b**) *p*-type region, under a forward bias. As the minority carriers diffuse further away from the edges of the depletion region, they recombine with majority carriers. The diffusion current of minority carriers is therefore reduced. But, this process also results in the flow of majority carriers in the opposite direction, which compensates the decrease in diffusion current with a drift current in the same proportion

carriers "consume" majority carriers (e.g., electrons). There thus must be a flow of majority carriers (e.g., electrons) in the opposite direction to resupply those lost in the recombination process. This flow of majority carriers generates a drift current.

Therefore, in the neutral regions, there are two components which make up the total electrical current: the diffusion current of minority carriers and the drift current of majority carriers. These are shown in Fig. 9.17 . This means, in particular, that there must be an electric field present in the neutral regions; otherwise there would not be any drift current. This apparently contradicts our assumption at the beginning of Subsect. 9.3.1 that there was no potential drop within the neutral regions. In fact, the potential drop is very small in comparison with any applied external bias voltage and therefore can be neglected in our model.

An analytical expression for the drift current can be easily determined, on each side of the *p-n* junction. Indeed, the total hole and electron current densities must be constant at the values given by the diode equation in Eq. (9.47). As we know the expression for the diffusion current densities $J_{\mathrm{h}}^{\mathrm{diff}}(x_1)$ and $J_{\mathrm{e}}^{\mathrm{diff}}(x_2)$ from Eq. (9.46), the drift current densities will be the difference:

$$\begin{cases} J_{\mathrm{h}}^{\mathrm{drift}}(x_2) = J_{\mathrm{e}}^{\mathrm{diff}}(0) - J_{\mathrm{e}}^{\mathrm{diff}}(x_2) \\ J_{\mathrm{e}}^{\mathrm{drift}}(x_1) = J_{\mathrm{h}}^{\mathrm{diff}}(0) - J_{\mathrm{h}}^{\mathrm{diff}}(x_1) \end{cases} \tag{9.57}$$

Recalling Eqs. (9.46) and (9.49), we get successively:

$$\begin{cases} J_{\mathrm{h}}^{\mathrm{drift}}(x_2) = -q\dfrac{D_{\mathrm{n}}}{L_{\mathrm{n}}}\Delta n_{\mathrm{p}} + q\dfrac{D_{\mathrm{n}}}{L_{\mathrm{n}}}\Delta n_{\mathrm{p}} e^{-\frac{x_2}{L_{\mathrm{n}}}} \\ J_{\mathrm{e}}^{\mathrm{drift}}(x_1) = q\dfrac{D_{\mathrm{p}}}{L_{\mathrm{p}}}\Delta p_{\mathrm{n}} - q\dfrac{D_{\mathrm{p}}}{L_{\mathrm{p}}}\Delta p_{\mathrm{n}} e^{-\frac{x_1}{L_{\mathrm{p}}}} \end{cases}$$

$$\begin{cases} J_{\mathrm{h}}^{\mathrm{drift}}(x_2) = q\dfrac{D_{\mathrm{n}}}{L_{\mathrm{n}}}\Delta n_{\mathrm{p}}\left(e^{-\frac{x_2}{L_{\mathrm{n}}}} - 1\right) \\ J_{\mathrm{e}}^{\mathrm{drift}}(x_1) = q\dfrac{D_{\mathrm{p}}}{L_{\mathrm{p}}}\Delta p_{\mathrm{n}}\left(1 - e^{-\frac{x_1}{L_{\mathrm{p}}}}\right) \end{cases} \tag{9.58}$$

It is important to remember that the sign convention chosen for $J_{\mathrm{h}}^{\mathrm{drift}}(x_2)$ is opposite that of axis x.

9.4 Deviations from the Ideal *p-n* Diode Case

Before deriving the ideal diode equation in the previous section, it was necessary to make several assumptions. In reality, these assumptions are not necessarily valid, and the ideal diode equation gives only qualitative agreement with actual measurements of the *I-V* characteristics of real *p-n* junction diodes. This deviation from the ideal case is mainly due to (a) generation of carriers in the depletion region, (b) surface leakage effects at the periphery of a real junction, (c) recombination of carriers in the depletion region, (d) the high-injection condition (when the injection of minority carriers exceeds the doping density), and finally (e) all the applied bias not being dropped across the depletion region due to series resistance effects. The above deviations are illustrated in the figure below. The special case of reverse breakdown will be discussed in Subsect. 9.4.3 (Fig. 9.18).

9.4.1 Reverse Bias Deviations from the Ideal Case

Part of the deviation of the leakage current from the ideal reverse saturation current arises from the thermal generation of electron-hole pairs within the space charge region. The built-in electric field separates these carriers and they drift toward the neutral regions of the diode. This drift results in an excess current that is in addition to the diffusion of minority carriers, discussed in the ideal case. Section 8.6 introduced the concept of thermal generation of carriers, and along with it a thermal generation rate per unit volume $G_{\mathrm{t}}(T)$, expressed in $\mathrm{cm}^{-3}\cdot\mathrm{s}^{-1}$. Since the volume of the depletion region is equal to WA, assuming no recombination occurs, the current due to generation in the depletion region can be expressed as:

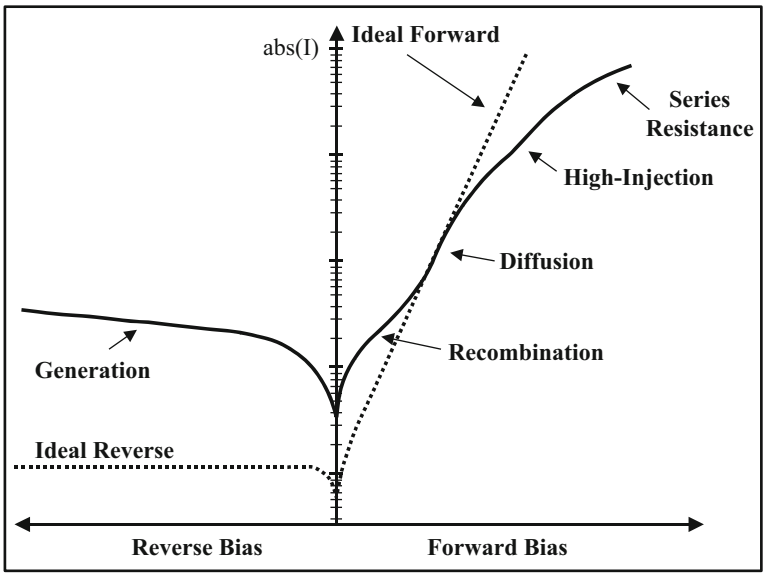


Fig. 9.18 The current-voltage characteristic for a real Si *p-n* junction diode (solid) does not exactly match the behavior of a Si junction diode predicted by the ideal diode model (dotted), both shown above in semilog scale. A real Si diode shows the following deviations from the ideal (diffusion limited) case: reverse leakage current due to thermal generation and surface leakage effects, recombination in the depletion region, high-injection deviation, and series resistance effects

$$I_{\text{gen}} = qWAG_{\text{t}}(T) \tag{9.59}$$

Under reverse bias the current can then be expressed as the sum of the diffusion and generation components:

$$I_{\text{rev}} = qA\left(\frac{D_{\text{p}}}{L_{\text{p}}}p_{\text{n}} + \frac{D_{\text{n}}}{L_{\text{n}}}n_{\text{p}}\right) + qWAG_{\text{t}}(T). \tag{9.60}$$

Since the depletion layer width (W) depends upon the applied bias, the reverse current of the diode now shows a bias dependence: as the reverse bias is increased, the depletion width widens, and hence this increases the generation current leading to a corresponding increase in the reverse leakage current as a function of applied bias. In addition to excess carriers arising from thermal generation, it is possible for external photoexcitation to create carriers in the depletion region – this is the case of a photodiode.

This leakage current is further compounded by the surface leakage. Surface leakage effects are due to the finite extent of the *p-n* junction area and the characteristics of the junctions that occur at the periphery of the diode. This is due primarily to ionic charges on or outside the semiconductor that induce corresponding image charges within the semiconductor. These charges create their own surface

depletion region that acts as a parallel conduction channel that bypasses the *p-n* junction and allows current to flow along the surface of the diode. Typically this leakage current increases with reverse bias.

9.4.2 Forward Bias Deviations from the Ideal Case

Under forward bias recombination dominates over the generation processes. In order to supply the carriers lost to recombination, the net external current flowing through the diode is increased. This current is called the recombination current (I_{rec}). The recombination rate is at its maximum near the center of the depletion region, where nearly equal number of electrons and holes are available to contribute to recombination. Assuming a linear variation of the potential across the depletion region, the potential at the center can be taken as $\frac{V_0-V}{2}$. In this case the carrier concentration at the center of the depletion region depends upon $\exp\left(-\frac{q(V_0-V)}{2k_bT}\right)$ rather than $\exp\left(\frac{q(V_0-V)}{k_bT}\right)$. The rate at which electrons and holes are recombining is then proportional to $\exp\left(\frac{qV}{2k_bT}\right)$. By introducing a material constant (I_{R0}) dependent upon the minority carrier recombination lifetimes in the respective halves of the depletion layer, and the overall depletion layer width, it becomes possible to arrive at an expression for the recombination current (I_R):

$$I_R \approx I_{R0}\exp\left(\frac{qV}{2k_bT}\right) \tag{9.61}$$

Combining this new equation for the recombination current together with the existing minority carrier diffusion current yields a new expression for the total current though the diode:

$$I = I_0\exp\left(\frac{qV}{k_bT}\right) + I_{R0}\exp\left(\frac{qV}{2k_bT}\right) \tag{9.62}$$

In working with real diodes, this equation is generally represented in an empirical form by introducing a new factor n called the ideality factor:

$$I \approx I_0\exp\left(\frac{qV}{nk_bT}\right) \tag{9.63}$$

In this combined equation, the ideality factor n tends toward 2 when recombination current dominates and tends toward 1 when diffusion current dominates and varies from 1 to 2 when both currents are comparable. In the case of silicon diodes operating at room temperature, both processes can be seen to operate as the current injection is increased from low to moderate levels.

Under higher levels of current injection (under forward bias), the diode enters the high-injection regime where the injected minority carrier density becomes

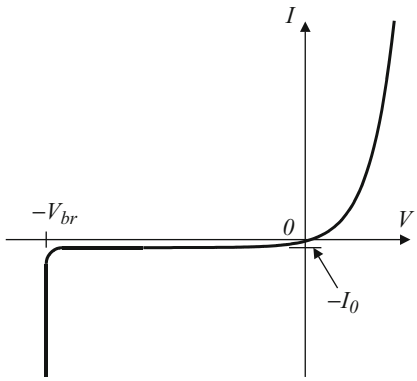


Fig. 9.19 Current-voltage characteristic for an ideal *p-n* junction diode showing a reverse breakdown. When the voltage across the *p-n* junction is equal to the reverse breakdown voltage, the current increases dramatically. If it is not limited, this current can damage the diode through heating

comparable or greater than the majority carrier density. In this case the current becomes proportional to $\exp\left(\frac{qV}{2k_bT}\right)$, as is shown in Fig. 9.19.

Under higher reverse bias, the contact potentials and the potential drop across the bulk regions of the semiconductor cease to be negligible, and the series resistance of the *p-n* diode no longer dominates. At this point the exponential increase in current begins to subside in favor of a more linear increase, limited by the series resistance of the diode. The empirical diode equation introduced above can be modified to take this behavior into account, by introducing a term (R_S) for the series resistance. Thus the equation becomes:

$$I \approx I_0 \exp\left(\frac{q(V - IR_S)}{nk_bT}\right) \tag{9.64}$$

9.4.3 Reverse Breakdown

In the ideal *p-n* junction diode model, we saw that the current through a *p-n* junction diode was limited by the saturation current $-I_0$ when a reverse bias was applied. Even in the non-ideal case the reverse current was seen to increase slowly. In reality, this model holds only up to a certain value of reverse bias $-V_{br}$, called the breakdown voltage. At that point, the current suddenly increases dramatically as shown in Fig. 9.19 . This phenomenon is called reverse breakdown. The peak value for the internal electric field strength (i.e., at $x = 0$) corresponding to this applied reverse bias is called the critical electric field.

This situation is not necessarily a damaging one for the *p-n* junction and is reversible, as long as the current can be limited to prevent too much power from being dissipated inside the device. Otherwise, parts of the device can be physically destroyed (e.g., melted).

There are two major mechanisms for the reverse breakdown: avalanche breakdown which occurs at higher reverse biases as a result of impact ionization and Zener breakdown which occurs at lower reverse biases as a result of tunneling across the junction.

9.4.4 Avalanche Breakdown

As a stronger reverse bias is applied, the electric field strength across the space charge region increases. The charge carrier particles, holes, and electrons which drift across the depletion region can therefore achieve higher velocities.

When the reverse bias is strong enough, typically higher than $6E_g/q$ and can even go up to 1000 V, the electric field strength can become so large that a hole or an electron can gain sufficient kinetic energy to impact on a semiconductor lattice atom and ionize it, or even break a chemical bond. This phenomenon is called impact ionization. It may seem conceptually difficult to envision a hole impacting on the crystal lattice, but this can be better understood when we realize that when a hole moves in one direction, it in fact corresponds to the motion of an electron in the opposite direction with the same velocity. An accelerated particle must typically acquire energy at least equal to the bandgap energy E_g in order to break a chemical bond, because this corresponds to the energy required to excite an electron from the valence band to the conduction band. Therefore, for wider bandgap semiconductors, higher electric field strength is necessary to ensure impact ionization.

As a result of impact ionization, an electron-hole pair (EHP) is created within the space charge region in addition to the impacting particle. The electron and the hole from the pair will then be spatially separated by the electric field present at that location: the electron drifting toward the *n*-type side and the hole toward the *p*-type side, as illustrated in Fig. 9.20.

The electrons and holes thus generated can themselves be further accelerated by the electric field. If they reach a sufficient high kinetic energy within the space charge region, they can in turn contribute to create additional EHPs through ionizing collisions. This results in a cascade or avalanche effect. One initial charge carrier thus has the potential to create many additional carriers, and a dramatic increase in current is achieved as the one shown in Fig. 9.19.

It is possible to characterize the avalanche breakdown quantitatively by introducing a multiplication factor M such that the reverse current near breakdown is given by MI_0 where I_0 is the saturation current. This factor actually means that an incident electron results in a total of M electron-hole pairs. This factor is empirically given by:

$$M = \frac{1}{1 - \left(\frac{V_r}{V_{br}}\right)^n} \tag{9.65}$$

where V_r is the reverse bias, V_{br} is the breakdown voltage, and n is an exponent in the range 3~6. From this expression, we clearly see that the reverse current, MI_0, increases sharply when V_r nears V_{br} as depicted in Fig. 9.19.

The avalanche process is more likely to occur when a wide enough space charge region can be sustained to ensure sufficient acceleration. This can be more easily achieved by using lightly doped *p-n* junctions because, if heavily doped junctions are used, another phenomenon can more easily occur: the tunneling of charge carriers from one side of the junction to the other.

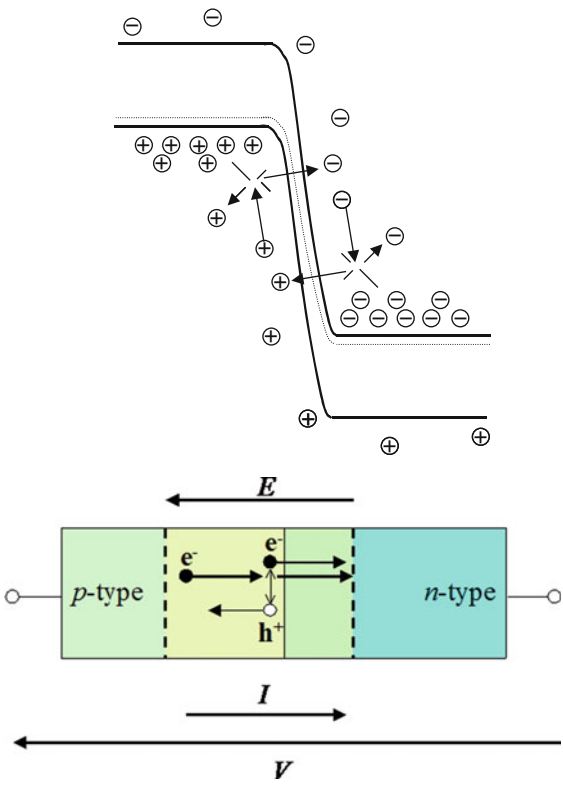


Fig. 9.20 Impact ionization process: under strong reverse bias, electrons and holes are injected into the depletion region; when they gain enough kinetic energy, they impact on the semiconductor lattice to create electron-hole pairs. These newly created carriers can then lead to the same impact ionization process if they can gain enough kinetic energy within the space charge region

Example

Q A voltage-stabilizing diode takes advantage of the steep slope in the breakdown regime to clamp the voltage. For such a kind of diode with $V_{\mathrm{br}} = -14$ V, estimate how many times the current will increase when the reverse bias goes from -13.990 to -13.995 V. Assume $n = 6$.

A The multiplication factor if given by: $M = \frac{1}{1-\left(\frac{V_{\mathrm{r}}}{V_{\mathrm{br}}}\right)^n}$. For the two reverse biases mentioned, we get the ratio of the multiplication factor:

$$\frac{M_1}{M_2} = \frac{1-\left(\frac{V_2}{V_{\mathrm{br}}}\right)^n}{1-\left(\frac{V_1}{V_{\mathrm{br}}}\right)^n} = \frac{1-\left(\frac{13.990}{14}\right)^6}{1-\left(\frac{13.995}{14}\right)^6} = 2$$

The current will thus increase by a factor 2 when the voltage is reduced by 0.005 V.

9.4.5 Zener Breakdown

Under a more moderate reverse bias, typically less than $6E_g/q$, the top of the valence band in the *p*-type side E_{Vp} is already higher than the bottom of the conduction band in the *n*-type side E_{Vc}. This situation is illustrated in Fig. 9.21. This means that the electrons at the top of the valence band in the *p*-type side have the same or higher energy than the empty states available at the bottom of the conduction band in the *n*-type side.

This staggering of the energy bands also results in a reduced spatial separation between the conduction and valence bands, as shown by *d* in Fig. 9.21 .Moreover, in heavily doped *p-n* junctions, the space charge region is already narrow (with a width *W*) and does not expand much under a moderate reverse bias.

The staggered alignment of the energy bands and their spatial proximity favor the tunneling of electrons from the valence band in the *p*-type side into the conduction band in the *n*-type side, as shown in Fig. 9.21. This leads to a negative current. This process is called the Zener effect. As there are many electrons in the valence band and many empty available states in the conduction band, the tunneling current can be substantial.

The Zener tunneling probability T_Z is strongly field dependent on the applied bias *V* and the bandgap E_g. It can be written as:

$$T_Z = \exp\left\{-\frac{4\sqrt{2m^*}}{3qV\hbar}E_g^{3/2}\right\} \tag{9.66}$$

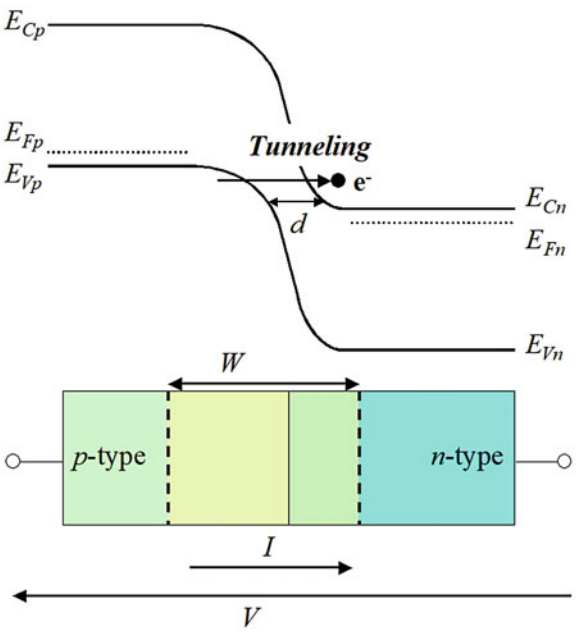


Fig. 9.21 Zener breakdown mechanism involving electrons tunneling from the valence band of the *p*-type side to the conduction band of the *n*-type side

9.5 Metal-Semiconductor Junctions

As we have already mentioned in Subsect. 9.2.6 and illustrated in the case of a *p-n* junction, two dissimilar materials in contact with each other and under thermal equilibrium must have the same value of Fermi energy.

When a metal is brought into contact with a semiconductor, a certain amount of band bending occurs to compensate the difference between the Fermi energies of the metal and that of the semiconductor. In fact, this difference in Fermi energy means that electrons in one material have a higher energy than in the other. These will therefore tend to flow from the former to the later material. There is thus a transfer of electrons across the metal-semiconductor junction in a similar way as the charge transfer in the case of a *p-n* junction. Such a junction is also often called a metallurgic junction or a metal contact because metals are commonly used in semiconductor industry to connect or "contact" a semiconductor material to an external electrical circuit.

The charge transfer can be readily achieved because, as we saw in Fig. 5.11 in Subsect. 5.2.7, the Fermi energy in a metal lies within an energy band, which makes it easy for electrons to be emitted from or received by a metal. This charge redistribution gives rise to a local built-in electric field which counterbalances this redistribution. When sufficiently large electric field strength is established around the metallurgic junction, the redistribution stops.

Since the overall charge neutrality must be maintained, the excess electrical charges inside the semiconductor and that inside the metal must be of an equal amount but with opposite signs. However, because a metal has a much higher charge density than a semiconductor, the width over which these excess charges spread inside the metal is negligibly thin in comparison to the width inside the semiconductor. This is somewhat similar to the case of a *p-n* junction with one side heavily doped. As a result, the built-in electric field and the band bending are primarily present inside the semiconductor as well. The following section aims at giving a quantitative description of the physical properties of a metal-semiconductor junction.

9.5.1 Formalism

The physical parameters which need to be considered in this description are depicted in Fig. 9.22. For the metal, these include its Fermi energy E_{Fm} and work function $\Phi_m > 0$. As we saw when discussing the photoelectric effect in Chap. 4, the work function of a metal is the energy required to extract one electron from the metal surface and pull it into the vacuum. In a more quantitative manner, the work function is the energy difference between the Fermi energy and the vacuum level as shown in Fig. 9.22. For the semiconductor, the parameters of interest also include its Fermi energy E_{Fs}, its work function $\Phi_s > 0$, and also its electron affinity $\chi > 0$. The latter is the energy required to extract one electron from the conduction band of the semiconductor into the vacuum and is given by the energy difference between the bottom

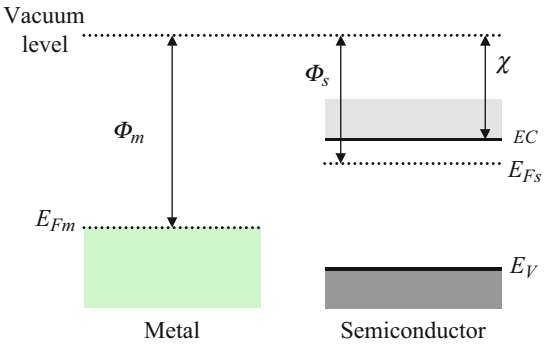


Fig. 9.22 Fermi energies, work functions in a metal and a semiconductor, when considered isolated from each other. The vacuum level is the same for both materials, but the Fermi energies are generally different

of the conduction band and the vacuum level. A few values of electron affinity for elements in the periodic table are given in Fig. A.12 in Appendix A.3.

The amount of band bending and the direction of electron transfer depend on the difference between the work functions of the metal and the semiconductor. When these materials are isolated, their vacuum levels are the same, as illustrated in Fig. 9.22. But, when these materials come into contact, the Fermi energy must be equal on both sides of the junction. The vacuum level is at an energy Φ_m above the top of the metal Fermi energy, while it is Φ_s above the semiconductor Fermi energy. This means that the energy bands in the semiconductor must shift upward by an amount equal to $\Phi_m - \Phi_s$ in order to align the Fermi energy on both sides of the junction.

On the one hand, if $\Phi_m > \Phi_s$, the energy bands of the semiconductor actually shift downward with respect to those of the metal, and electrons are transferred from the semiconductor into the metal, as shown in Fig. 9.23. The signs of the charge carriers which appear on either side of the junction and the direction of the built-in electric field, also shown in Fig. 9.23, are determined from the analysis conducted for a *p-n* junction. On the other hand, if $\Phi_m < \Phi_s$, the energy bands in the semiconductor shift upward with respect to those of the metal, and the electrons are transferred from the metal into the semiconductor.

9.5.2 Schottky and Ohmic Contacts

The electrical properties of a metal-semiconductor junction depend on whether a depletion region is created as a result of the charge redistribution. This phenomenon in turn depends on the difference in work function Φ_m-Φ_s, and on the type of the semiconductor (*n*-type or *p*-type).

Indeed, we know that when $\Phi_m > \Phi_s$, electrons are extracted from the semiconductor into the metal.

If the semiconductor is *n*-type, then this process depletes the semiconductor of its electrons or majority charge carriers. A depletion region thus appears near the junction, and we obtain a diode-like behavior similar to a *p-n* junction when an

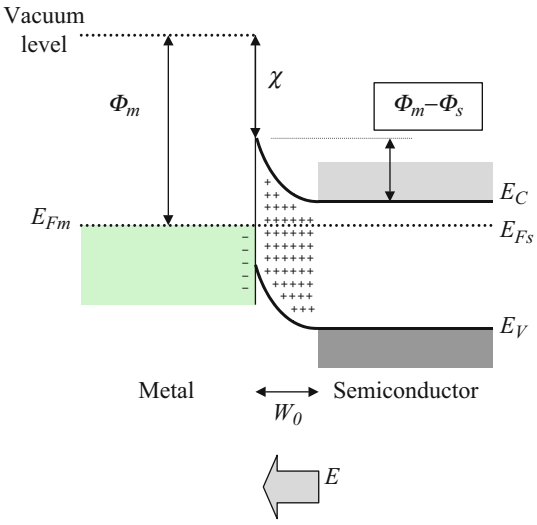


Fig. 9.23 Energy levels, accumulated charge carriers, and built-in electric field in a metal-semiconductor junction. When the metal and the semiconductor are brought into contact, at equilibrium, the energy band profile of the semiconductor near the junction is modified so that the Fermi energies become equal in both materials

external bias is applied. This is shown in Fig. 9.24a. This situation is often called a rectifying contact or Schottky contact.

However, if the semiconductor is *p*-type, the electrons which are extracted from the semiconductor are taken from the *p*-type dopants which then become ionized. This process thus creates more holes or majority charge carriers. In this case, there is no depletion region, but rather majority carriers are accumulated near the junction area, and we do not observe a diode-like behavior. Majority carriers are free to flow in either direction under the influence of an external bias. This is shown in Fig. 9.25a. This situation is often called an ohmic contact and the current-voltage characteristics are linear.

If we now consider $\Phi_m < \Phi_s$, electrons are extracted from the metal into the semiconductor. The previous analysis needs to be reversed. In other words, for an *n*-type semiconductor, the junction will be an ohmic contact, while for a *p*-type semiconductor, the junction will be a Schottky contact.

These four configurations are shown in Figs. 9.23 and 9.24 and summarized in Table 9.1.

In the case of a Schottky contact, the existence of the depletion region means that there is a potential barrier across the junction which can be shifted by an amount equal to $-qV$ when an external voltage V is applied between the metal and the semiconductor. This in turn influences the current flow in a similar way as for a *p-n* junction. This is shown in Fig. 9.26 for the case of an *n*-type semiconductor. It is however important to understand that majority carriers are responsible for the current transport in a metal-semiconductor junction, whereas in a *p-n* junction, it is due to the minority carriers.

The sign convention for a metal-semiconductor junction is the same as for a *p-n* junction by considering the type of the semiconductor. Although the current

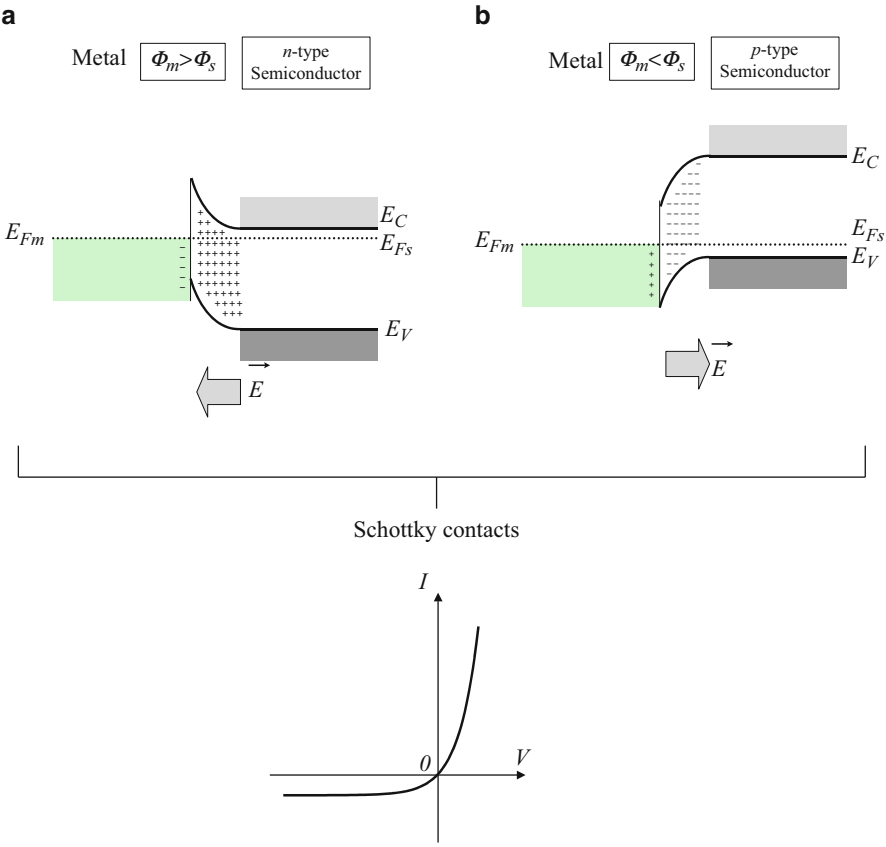


Fig. 9.24 These two of the four possible metal-semiconductor junction configurations lead to a Schottky contact: (**a**) $\Phi_m > \Phi_s$ and *n*-type, (**b**) $\Phi_m < \Phi_s$ and *p*-type. A Schottky contact is obtained in each case because the majority carriers in the semiconductor experience a potential barrier which prevents their free movement across the metal-semiconductor junction, and therefore as shown at the bottom of the figure, the *I-V* characteristic shows rectifying behavior

transport mechanism in a Schottky contact is somewhat different from that in a *p-n* junction, the current-voltage relation for an ideal Schottky contact has a similar expression as for an ideal *p-n* junction:

$$I = I_0\left(e^{\frac{qV}{k_b T}} - 1\right) \tag{9.67}$$

where I_0 is the reverse saturation current and is exponentially proportional to the difference between the metal work function Φ_m and the semiconductor electron affinity χ:

$$I_0 = AB_e T^2 e^{\left(-\frac{(\Phi_m - \chi)}{k_b T}\right)} \tag{9.68}$$

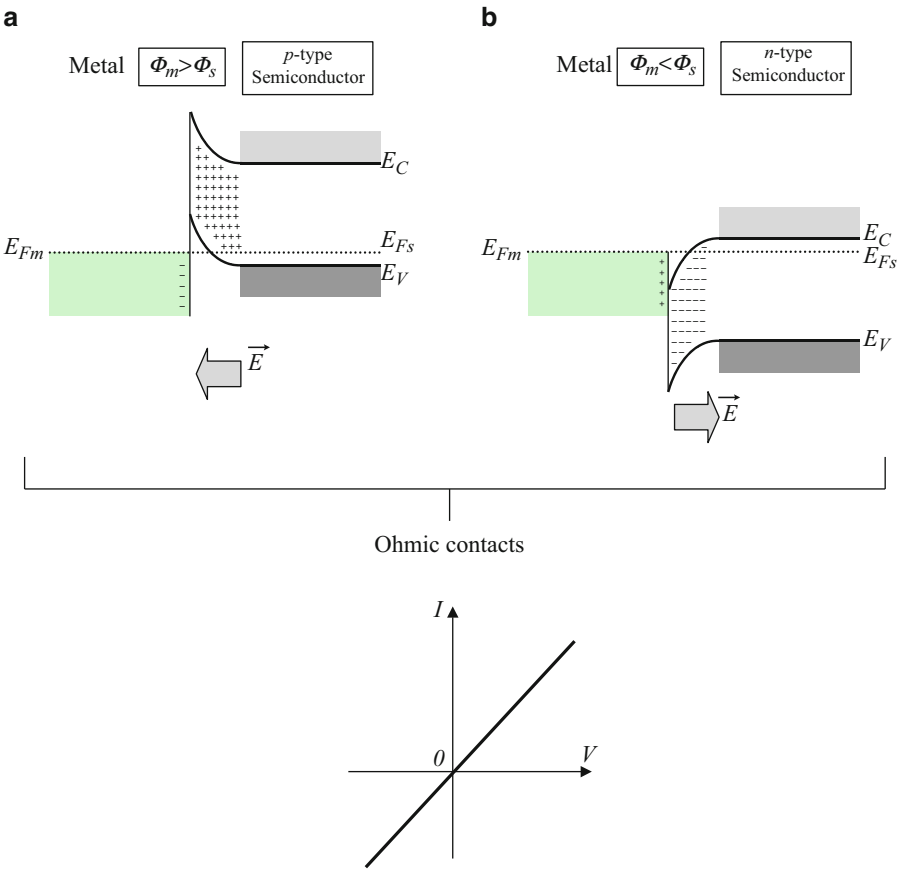


Fig. 9.25 These two of the four possible metal-semiconductor junction configurations lead to an ohmic contact: (**a**) $\Phi_m > \Phi_s$ and *p*-type, (**b**) $\Phi_m < \Phi_s$ and *n*-type. Unlike the configurations shown in Fig. 9.24, the energy band profiles here are such that the majority carriers in the semiconductor can move across the metal-semiconductor junction without experiencing a potential barrier, and therefore as shown at the bottom of the figure, the *I-V* characteristic shows ohmic behavior

Table 9.1 Four possible metal-semiconductor junction configurations and the resulting contact types

	Semiconductor	Junction
$\Phi_m > \Phi_s$	*n*-type	Schottky
$\Phi_m < \Phi_s$	*p*-type	Schottky
$\Phi_m > \Phi_s$	*p*-type	Ohmic
$\Phi_m < \Phi_s$	*n*-type	Ohmic

B_e is the effective Richardson constant, and for most metal-semiconductor Schottky junctions, it varies from 10 to 100 K^{-2} cm^{-2}. The quantity $(\Phi_m - \chi)$ is often denoted $q\Phi_B$, where Φ_B is called the Schottky potential barrier height. For a real Schottky contact, one needs to take into account thermionic emission (Appendix

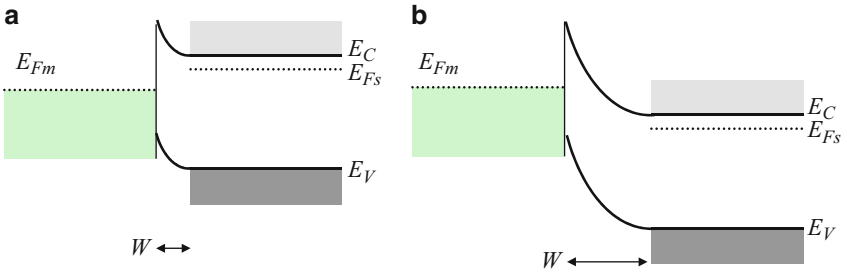


Fig. 9.26 Band alignment in a Schottky metal-(*n*-type) semiconductor contact under (**a**) forward bias where the potential barrier is reduced, and under (**b**) reverse bias where the potential barrier is increased, thus reducing the tunneling of carriers

A.9), as well as impurity and interface states. In this case, the current-voltage relation is given by:

$$I = I_0\left(e^{\frac{qV}{nk_bT}} - 1\right) \tag{9.69}$$

where n is the ideality factor as mentioned before and is typically between 1 and 2.

9.6 Summary

In this chapter, we have presented a complete mathematical model for an ideal *p-n* junction, based on an abrupt homojunction model and the depletion approximation. We introduced the concepts of a space charge region, built-in electric field, built-potential, and depletion width at equilibrium. We have discussed the balance of electrical charges, as well as that of the diffusion and drift currents within the space charge region.

The non-equilibrium properties of *p-n* junctions have also been discussed. The forward bias and reverse bias conditions were examined. We emphasized the importance of minority carrier injection and extraction. We derived the diode equation and understood the nature of the currents outside the space charge region. We have discussed the avalanche and Zener breakdown mechanisms as deviations from the ideal *p-n* junction diode behavior under strong reverse bias conditions.

Finally, we presented the electrical properties of metal-semiconductor junctions and introduced the concepts of Schottky and ohmic contacts.

Problems

1. A *p-n* junction diode has a concentration of $N_A = 10^{17}$ acceptor atoms per cm^3 on the *p*-type side and a concentration of N_D donor atoms per cm^3 on the *n*-type side. Determine the built-in potential V_0 at room temperature for a germanium diode for values of N_D ranging from 10^{14} to 10^{19} cm^{-3}. Also determine the peak

value of the electric field strength for this same range, and plot both of these values as a function of N_D on a semilog scale.

2. Consider a GaAs step junction with $N_A = 10^{17}$ cm^{-3} and $N_D = 5 \times 10^{15}$ cm^{-3}. Calculate the Fermi energy in the *p*-type and *n*-type regions at 300 K. Draw the energy band diagram for this junction. Determine the built-in potential from the diagram and from Eq. (9.22). Compare the results.
3. Consider an asymmetric p^+-*n* junction, which has a heavily doped *p*-type side relative to the *n*-type side, i.e., $N_A >> N_D$. Determine a simplified expression for the width of the space charge region given in Eq. (9.23).
4. Calculate the depletion width for a Si *p-n* junction that has been doped with 10^{18} acceptor atoms per cm^3 on the *p*-type side and 10^{16} donor atoms per cm^3 on the *n*-type side. Compare this depletion width to the width of the depletion region on the n-side (from Eq. (9.33)). What percentage of the width lies within the *n*-type semiconductor. T = 300 K.
5. A silicon *p-n* diode with $N_A = 10^{18}$ cm^{-3} has a built-in voltage of 0.814 eV and capacitance of 10^{-8} F·cm^{-2} at an applied voltage of 0.5 V. Determine the donor density. A = 1 cm^2.
6. Plot the diode equation for an ideal Si *p-n* junction diode with an area 50 μm^2, an acceptor concentration $N_A = 10^{18}$ cm^{-3}, a donor concentration $N_D = 10^{18}$ cm^{-3}, recombination lifetimes equal to $\tau_n = \tau_p = 1$ μs, and diffusion coefficients equal to $D_n = 35$ cm^2·s^{-1} and $D_p = 12.5$ cm^2·s^{-1}.
7. Consider a Si *p-n* step junction with $N_A = 10^{17}$ cm^{-3} and $N_D = 10^{16}$ cm^{-3}, with recombination lifetimes $\tau_p = 0.1$ μs and $\tau_n = 0.01$ μs and carrier mobilities $\mu_h = 450$ cm^2/Vs and $\mu_e = 800$ cm^2/Vs at 300 K.
8. Determine the total reverse saturation current density, the reverse saturation current density due to holes and that due to electrons.
9. Assume a forward bias equal to $V_0/2$ is applied, where V_0, the built-in potential, is equal to 0.7546 V. Calculate the injected minority carrier currents at the edges of the space charge region.
10. Assume a reverse bias equal to $-V_0/2$ is applied. Calculate the minority carrier currents at the edges of the space charge region.
11. A Si *p-n* junction is doped with an acceptor concentration $N_A = 5 \times 10^{18}$ cm^{-3} and a donor concentration $N_D = 5 \times 10^{15}$ cm^{-3}. The critical electric field strength for breakdown is equal to 10^5 V·cm^{-1}. Determine the breakdown voltage and the corresponding depletion width. Do the same for a donor concentration $N_D = 5 \times 10^{17}$ cm^{-3}.
12. Consider an ideal metal-semiconductor junction between *p*-type silicon and polycrystalline aluminum. The Si is doped with $N_A = 5 \times 10^{16}$ cm^{-3}. The metal work function is 4.28 eV and the Si electron affinity is 4.01 eV. Draw the equilibrium band diagram and determine the barrier height ϕ_B.
13. Consider the same silicon-aluminum metal-semiconductor junction. The cross-sectional area of the junction is 10 μm^2. Assume that B_e is 30 AK^{-2} cm^{-2} and the ideality factor *n* is 1. Calculate the reverse saturation current and plot the *I-V* curve as a function of applied bias.

Further Reading

Ashcroft NW, Mermin ND (1976) *Solid state physics*. Holt Rinehart and Winston, New York
Neudeck GW (1989) The PN junction diode. Addison-Wesley, Reading
Pierret RF (1989) Advanced semiconductor fundamentals. Addison-Wesley, Reading
Sapoval B, Hermann C (1995) Physics of semiconductors. Springer, New York
Streetman BG (1990) Solid state electronic devices. Prentice-Hall, Englewood Cliffs
Sze SM (1981) Physics of semiconductor devices. Wiley, New York
Wang S (1989) Fundamentals of semiconductor theory and device physics. Prentice-Hall, Englewood Cliffs

Optical Properties of Semiconductors 10

10.1 Introduction

In previous chapters, we introduced the reader to the fundamental concepts of quantum mechanics, band structure, and semiconductor physics. In this chapter we have the opportunity to apply this acquired knowledge of the electronic structure of solids to understand the optical properties. We do this by modeling the optical response properties, in particular the permittivity of the solid. We present the formalism which allows one to calculate the permittivity and then study how this permittivity affects the light penetrating the solid. We shall demonstrate how band structure and free electrons determine the permittivity, and therefore the way light propagates in a solid, and how much of this light gets absorbed. We shall investigate under what circumstances the lattice can couple to photons and how this coupling can affect the velocity of light in a medium. But we shall see in the next chapters that band structure depends on the dimensionality of the system, and we have already seen in Chaps. 8 and 9 that carriers can be added or neutralized in semiconductors. So it turns out that just in the same way that the energy bands can be engineered, so can the optical properties. Atom by atom growth and miniaturization are modern key engineering tools, but so is the application of external electric and magnetic fields. In the last sections of this chapter, we therefore investigate how an electric or a magnetic field modifies the band structure and how this reflects on the optical properties. The fundamental concepts developed in this chapter are a necessary prerequisite to understand the way optical methods can be used to characterize the electronic structure of semiconductors as is described in Chap. 15.

Maxwell showed many years ago that light is an electromagnetic wave which travels in space and in media and interacts with the medium because the electric field vector of the light can polarize the medium and move the free charges about and produce a time-dependent current. The field changes the medium which acts back on the wave, becomes the wave, and affects its speed and amplitude.

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_10

Quantum theorists, as we have seen in Chap. 4, have shown that electromagnetic waves can also be viewed as moving vibrations which consist of bundles of energy, as particles called photons, which each carry a specific quantum of energy proportional to the frequency of the vibration ν; the energy is $h\nu = \hbar\omega$ where ω is the angular frequency. In analogy to phonons, the quantum of lattice vibrations, it turns out in practice that for most purposes, the classical description of light (photons) is quite adequate, and we shall therefore continue our study of optical properties using Maxwell's equations. When necessary we will change to quantum mechanics, but throughout we shall also freely use the term photons to describe the particles which constitute a beam of light.

10.2 The Complex Refractive Index of a Solid

10.2.1 Maxwell's Equations

In order to understand how light interacts with a semiconductor, we need to say a few words about light propagation in a given medium. Consider a medium which has both bound electrons and free electrons. The propagation of light in this medium is described by Maxwell's equations. Maxwell's equations can be written in a form which from the very beginning distinguishes a conducting medium from a nonconducting medium, by writing:

$$\vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{10.1}$$

$$\vec{\nabla} \times \vec{H} = \sigma(\omega)\,\vec{E} + \frac{\partial \vec{D}}{\partial t} \tag{10.2}$$

$$\vec{\nabla} \cdot \vec{D} = \rho \tag{10.3}$$

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{10.4}$$

where $\sigma(\omega)$ is the complex frequency-dependent conductivity of the medium with a density of ρ mobile charges, and $\vec{E}$, $\vec{D}$, $\vec{H}$, and $\vec{B}$ are the electric field, displacement, magnetic field, and magnetic flux, respectively.

We are mainly interested in neutral media, so we shall put $\rho = 0$ and assume that the relative permittivity ε_b of a medium with bound charges in $\vec{D} = \varepsilon_0 \varepsilon_b \vec{E}$ is time independent and $\vec{D} = \varepsilon_0 \vec{E} + \vec{P}$ where $\vec{P}$ is the bound polarization vector which gives the electric dipole moment per unit volume and ε_0 is the permittivity of free space. We also assume that the medium is not magnetic so that $\vec{B} = \mu\mu_0 \vec{H}$, $\mu = 1$, the permeability of free space. Using the fact that the velocity of light in free

space is $c^2 = (\mu_0 \varepsilon_0)^{-1}$, one can combine Eqs. (10.1 and 10.2) by taking the "curl" (or "rot") of Eq. (10.1) to give the wave equation for an EM wave as:

$$\nabla^2 \vec{E} = \frac{1}{c^2}\left(\varepsilon_b \frac{\partial^2 \vec{E}}{\partial t^2} + \frac{\sigma}{\varepsilon_0}\frac{\partial \vec{E}}{\partial t}\right) \tag{10.5}$$

As we will see, this equation describes a traveling wave that can be solved by assuming that the electric field of the light is of the form:

$$\vec{E} = \vec{E}_0 \exp\left[i\left(\vec{k} \cdot \vec{r} - \omega t\right)\right] \tag{10.6}$$

The substitution of Eq. (10.6) into Eq. (10.5) then gives rise to the requirement that to be a solution; the length of the vector $\vec{k}$ (the wavevector) must satisfy the complex equation:

$$k = \frac{\omega}{c}\left(\varepsilon_b + \frac{i\sigma}{\varepsilon_0 \omega}\right)^{1/2} \tag{10.7}$$

since the wavevector $k = k_0 = \frac{\omega}{c}$ in free space, we can interpret the square root factor in Eq. (10.7) as the complex refractive index of the material $\bar{N}$:

$$\bar{N} = \left(\varepsilon_b + \frac{i\sigma}{\varepsilon_0 \omega}\right)^{1/2} \tag{10.8}$$

We recall that in this representation, ε_b refers to the (relative) bound electron permittivity and is itself normally a complex quantity. This is why some authors prefer to work with a total relative complex permittivity ($\varepsilon_t(\omega) = \varepsilon_b + \frac{i\sigma}{\varepsilon_0 \omega}$) and define $\vec{D} = \varepsilon_0 \varepsilon_t \vec{E}$, which includes both the complex free and the complex bound electron permittivities. In the notation that we have chosen, the conductivity of the medium is made explicit, and $\sigma(\omega)$ is the complex frequency-dependent conductivity of the system, the real part of which is the AC conductivity or, with geometry factor (area /length), the "conductance" of the system. The imaginary part then corresponds to ωC where C is the capacitance. Indeed if we separate the bound electron permittivity into real ε_r and imaginary parts ε_i, we have:

$$\bar{N}^2 = \varepsilon_r + i\left(\varepsilon_i + \frac{\sigma}{\varepsilon_0 \omega}\right) \tag{10.9}$$

The free electron contribution to the permittivity is now by definition:

$$\varepsilon_f(\omega) = i\frac{\sigma(\omega)}{\varepsilon_0 \omega} \tag{10.10}$$

We can now rewrite the complex refractive index and complex wavevector as:

$$\bar{N} = \bar{n} + i\kappa \tag{10.11}$$

$$k = \frac{\bar{n}\omega}{c} + \frac{i\kappa\omega}{c} = \bar{N}k_0 \tag{10.12}$$

The imaginary part of Eq. (10.11) acquires physical significance as soon as we substitute Eq. (10.12) back into the wave solution Eq. (10.6) and for simplicity assume propagation in the z-direction only, then we have:

$$\vec{E} = \vec{E}_0 \exp\left\{ i\omega\left(\frac{\bar{n}z}{c} - t\right)\right\}\exp\left(-\frac{\omega\kappa z}{c}\right) \tag{10.13}$$

For $\underset{0}{\vec{E}} = E_0^x \vec{x}$ the corresponding H_0^y is given by $H_0^y = \bar{N}\sqrt{\frac{\varepsilon_0}{\mu_0}}E_0^x$ where we also have from Eq. (10.9) and $\sigma = \sigma_r + i\sigma_i$:

$$\begin{aligned} \bar{n}^2 - \kappa^2 &= \varepsilon_r - \frac{\sigma_i}{\omega\varepsilon_0} \\ 2\bar{n}\kappa &= \varepsilon_i + \frac{\sigma_r}{\varepsilon_0\omega} \end{aligned} \tag{10.14}$$

The medium has modified the electromagnetic wave or photon, in two ways. It has changed the velocity of propagation from c to $c/\bar{n}$, and it has given rise to damping. The damping is due to the imaginary part of k and is caused by the absorption of electromagnetic energy in the medium. From Eq. (10.14) it follows that one principal source of absorption is the conductivity term. But loss of amplitude can also be caused by the bound electrons absorbing light energy and getting excited into higher-energy levels in the solid. Bound electron absorption happens at relatively high frequencies, so that in practice, as we shall see later, the low-frequency damping is mainly due to free charges, and the high-frequency damping is mainly due to band-to-band absorption. Noting that the energy density is proportional to the square of the electric field amplitude, we recover the Beer-Lambert law:

$$\begin{aligned} |E|^2 &= |E_0|^2 e^{-\alpha z} \\ \alpha &= 2\kappa\frac{\omega}{c} \end{aligned} \tag{10.15}$$

where α is the absorption coefficient and measured in units of m^{-1} in the MKS units as used here.

A word of caution as to the definition of the absorption coefficient. In the transmission of light through a material, the electric field amplitude can decay not just because of absorption. The decay may be due to a disorder, i.e., scattering, and this is why some authors prefer to compute the power dissipated per unit length.

The optical power density of the electromagnetic wave in units of W/m^2 is given by the time averaged Poynting vector:

$$\vec{S} = 1/2\mathrm{Re}\left(\vec{E} \times \vec{H}^*\right) = \frac{\bar{n}c}{2}\varepsilon_0 \left(E_0^x\right)^2 e^{-\alpha z}\,\vec{z} \tag{10.16}$$

10.2.2 Reflectivity

Before getting on with the evaluation of the complex permittivities and conductivity, it is convenient to investigate what happens when photons, or in other words the light beam, are incident onto a medium with complex refractive index coming from free space. Consider for simplicity normal incidence as shown in Fig. 10.1.

The wavevector $k = k_{0z}$ has a z-component only and is traveling in the z-direction. We assume that the wave is polarized with its E_x vector lying in the x-y plane and pointing in the x-direction. The boundary of the two media is at $z = 0$, so in the region $z > 0$, i.e., in the medium, the EM wave is traveling in one direction only and given by:

$$E_x(t,z) = E_0 \exp\left(i\omega\frac{\bar{N}z}{c} - t\right) \tag{10.17}$$

We are assuming that the medium is thick, so that there is no back reflected wave from a second interface. In the $z < 0$ region, free space, we have both the incoming wave E_i and the reflected wave E_r:

$$E_x(t,z) = E_i \exp\left[i\omega\left(\frac{z}{c} - t\right)\right] + E_r \exp\left[-i\omega\left(\frac{z}{c} + t\right)\right] \tag{10.18}$$

The continuity requirement of the electric field at the boundary $z = 0$ gives us:

$$E_0 = E_i + E_r \tag{10.19}$$

Knowing the electric field allows us to deduce the magnetic field using Maxwell's equation so that, for $z > 0$:

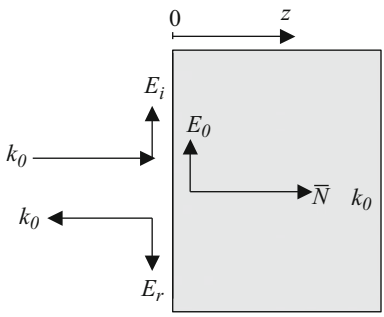


Fig. 10.1 The reflection and transmission process expressed in terms of a diagram

$$H_y = \frac{-1}{\omega\mu_0}\left(\bar{N}k_0\right)E_x \tag{10.20}$$

and then use the continuity condition for H at the boundary, which gives:

$$\bar{N}E_0 = E_i - E_r \tag{10.21}$$

Note the magnetic field at $z = 0$ depends on the direction of propagation.
From this pair of equations, we can deduce the relation:

$$\frac{E_r}{E_i} = \frac{1-\bar{N}}{1+\bar{N}} \tag{10.22}$$

The ratio of reflected to incident power is the reflectivity $R = \left|\frac{E_r}{E_i}\right|^2$ of the medium, and the squared of the absolute value of Eq. (10.22) gives:

$$R = \left|\frac{1-\bar{N}}{1+\bar{N}}\right|^2 = \frac{\left(\bar{n}-1\right)^2+\kappa^2}{\left(\bar{n}+1\right)^2+\kappa^2} \tag{10.23}$$

Thus knowing the complex refractive index as a function of frequency allows us to immediately calculate the reflectivity of a medium. One should note that there is, at this stage, no simple intuitive way of seeing from Eq. (10.23) when a medium is highly reflective or not. One has to calculate the equation. In order to develop this intuition, we need to go one step forward and actually derive explicit expressions for the refractive index in limiting situations of interest. Before that, it is useful and instructive to also consider the optical transmission and reflection through a slab of finite thickness d.

10.2.3 Transmission Through a Thin Slab

If R is the reflectivity, A the absorbance, and T the transmissivity, for a slab of finite thickness d, we must, by energy conservation, have$R + T + A = 1$. In the region $z < 0$, we have two waves as before, the incoming and reflected waves E_i and E_{r1}. In region $z > 0$, inside the medium, the EM wave now also consists of two components, one moving forward as before E_{t1} and one back reflected from the second interface E_{r2}. The second interface is at $z = d$. The waves E_{t1} and E_{r2} are traveling inside the medium and are therefore simply related via Eq. (10.6) to the corresponding waves at $z = d$, E'_{t1}, and E'_{r2} by a phase factor $e^{\pm dk_0 N}$. Outside, we have the outgoing transmitted wave into free space E_{t2}. The boundary condition for the electric and magnetic field must be taken at $z = 0$ and at $z = d$ and give four equations for four unknowns (E_{r1}, E_{t1}, E_{r2}, E_{t2}) and allow an explicit solution of this problem as before. The transmissivity T defined as $\left|\frac{E_{t2}}{E_i}\right|^2$ becomes:

$$T = \frac{\left(1 - |r_{01}|^2\right)^2 e^{-\alpha d}}{\left(1 - |r_{01}|^2 e^{-\alpha d}\right)^2} \qquad (10.24)$$
$$|r_{01}|^2 = \left|\frac{1-\bar{N}}{1+\bar{N}}\right|^2$$

where $\alpha = 2\frac{\omega}{c}\kappa$ is the absorption coefficient in the medium and $|r_{01}|^2$ can be recognized to be from Eq. (10.23) the reflectivity of the slab if it were very thick. The reflectivity of the slab R is given by the ratio $\left|\frac{E_{r1}}{E_i}\right|^2$ and correspondingly:

$$R = \frac{|r_{01}|^2\left(1 - e^{-\alpha d}\right)^2}{\left(1 - |r_{01}|^2 e^{-\alpha d}\right)^2} \qquad (10.25)$$

From Eqs. (10.24 and 10.25), one can now deduce the absorbance $A - 1 - R - T$. In the limit of a very thick slab, $e^{-\alpha d} \to 0$and R reduces to the previous expression.

10.3 The Free Carrier Contribution to the Complex Refractive Index

10.3.1 The Drude Theory of Conductivity

In Chap. 7 we calculated the conductivity of a nearly free electron gas in a *dc* field using a very simple relaxation time model also called the Drude model. We now consider the same model but allow the electric field to be time dependent. In particular, this can be the electric field vector of an impinging light (EM) wave as considered above.

Newton's law for carriers of effective mass m^* in a time-dependent field $E_0 e^{-i\omega t}$ and subject to the frictional force (Chap. 8) can be written as:

$$m^* \frac{d^2x}{dt^2} + m^* \frac{dx}{dt}\frac{1}{\tau} - qE(t) \qquad (10.26)$$

The displacement $x(t)$ of the particle is also expected to oscillate in time and follow the field, so that a solution to this equation could be $x(t) = x_0 e^{-i\omega t}$. Substitute this trial function into Eq. (10.26) and differentiate in time. The condition that this can be a solution to Eq. (10.26) is that:

$$-m^*\omega^2 x_0 - m^* \frac{i\omega}{\tau} x_0 = -qE_0 \qquad (10.27)$$

which immediately allows us to extract the amplitude x_0 as:

$$x_0 = \frac{q\tau}{m^* i\omega}\left(\frac{1}{1 - i\omega\tau}\right)E_0 \tag{10.28}$$

When negative charges move against a positive background, they produce a dipole. The polarization density produced by the time-varying field is the next quantity of interest. Thus the polarization density produced by a density n_c of displaced electronic charges is given by:

$$P = -n_c q x(t) = -\frac{n_c q^2 \tau}{m^* i\omega}\left(\frac{1}{1 - i\omega\tau}\right)E_0 e^{-i\omega t} \tag{10.29}$$

from which we can now also deduce the polarizability or optical susceptibility as the ratio:

$$\alpha_p(\omega) = \frac{P_c(t)}{E_0 e^{-\omega t}} \tag{10.30}$$

and write:

$$\alpha_p(\omega) = -\frac{n_c q^2 \tau}{m^* i\omega}\left(\frac{1}{1 - i\omega\tau}\right) \tag{10.31}$$

And for the complex conductivity we have, from the current:

$$-\frac{n_c q\frac{dx}{dt}}{E_0 e^{-i\omega t}} = \sigma(\omega) = \frac{n_c q^2 \tau}{m^*}\left(\frac{1}{1 - i\omega\tau}\right) \tag{10.32}$$

From the polarizability, we can deduce the relative permittivity produced by nearly free electrons, in the usual electrodynamic way ($\varepsilon_f = \left(1 + \frac{\alpha_p}{\varepsilon_0}\right)$):

$$\varepsilon_f = 1 - \frac{n_c q^2 \tau}{\varepsilon_0 m^* i\omega}\left(\frac{1}{1 - i\omega\tau}\right) \tag{10.33}$$

It is convenient and useful to rewrite the relative permittivity in a form which involves the plasma frequency ω_p and rewrite it as:

$$\varepsilon_f = 1 - \frac{\omega_p^2}{\omega^2}\left(\frac{\omega\tau(\omega\tau - i)}{1 + (\omega\tau)^2}\right) \tag{10.34}$$

$$\omega_p^2 = \frac{n_c q^2}{m^* \varepsilon_0} \tag{10.35}$$

The plasma frequency is the frequency at which the electron gas would oscillate as a whole if the electrons were collectively displaced and released from their equilibrium position. This can happen as follows: the electrons (n_c per unit volume)

are all displaced by a field by a distance x. This displacement causes a polarization $P = n_c q x$, which produces an electric field and restoring force$=-n_c q^2 x/\varepsilon_0$. The restoring force acting on each electron is proportional to the displacement, and we thus have simple harmonic motion with frequency $\omega_p = \sqrt{\frac{n_c q^2}{m^* \varepsilon_0}}$.

Now that we have the permittivity, we can apply it to find out a bit more about the optical properties of systems with free charge: metallic systems. Assume that the solid in question is a pure nearly free electron gas embedded in a jellium. A real metal will have both free and bound electron contributions, but the free electron responds strongly, and this term is often dominant. We will consider the bound electrons in the next section. There are two interesting limits for the refractive index.

First, when $\omega\tau << 1$, the second complex term on the right-hand side of Eq. (10.34) dominates and $\varepsilon_f(\omega)$ reduces to:

$$\varepsilon_f(\omega) \sim i \frac{n_c q^2 \tau}{\varepsilon_0 m^* \omega} \tag{10.36}$$

the permittivity is purely imaginary, and the square root of i has an equal real and imaginary part of $\cos(\pi/4)$ and $\sin(\pi/4)$, giving:

$$n(\omega) = \left\{ \frac{n_c q^2 \tau}{2\varepsilon_0 m^* \omega} \right\}^{1/2} \tag{10.37}$$

and which via Eq. (10.27) gives rise to a high reflection coefficient for small frequencies.

Secondly, in the limit that $\omega\tau \gg 1$, the relative permittivity is dominated by the real part and reduces to the form:

$$\varepsilon_f(\omega) \sim \left\{ 1 - \frac{\omega_p^2}{\omega^2} \right\} \tag{10.38}$$

In this limit the permittivity is purely real, which means that there is no absorption. It is also negative when the frequency is smaller than the plasma frequency. This implies that in this region, the refractive index is purely imaginary, and according to Eq. (10.23), we have perfect reflectance. Perfect reflectance means that the wave is not allowed to travel inside the medium. It can just tunnel in a little and go back out again. The fact that the permittivity can become less than 1, and even negative, turns out to be one of the most significant properties of metallic systems. It gives rise to the phenomenon of surface plasmon excitations at metal-dielectric interfaces and in metal particles. These are collective charge oscillations which can be excited by light, are mobile, and absorb the light very efficiently when the energy momentum conservation laws for their production are satisfied. Indeed when $\varepsilon(\omega) = 0$, a transverse wave can excite a longitudinal wave. The topic of surface plasmons is outside the scope of this textbook, but the reader can consult the textbook by Peyghambarian et al. (1993).

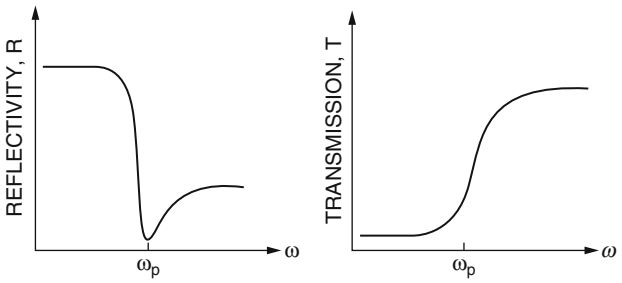


Fig. 10.2 The reflectivity and transmissivity of an electron gas (thin film) (Peyghambarian et al. 1993, p. 62, Fig. 3., Reprinted with permission)

When the frequency is above the plasma frequency, the permittivity is real and when $\varepsilon < 1$, it vanishes at the plasma frequency. The refractive index in this limit becomes:

$$\bar{n}(\omega) = \sqrt{\left(1 - \frac{\omega_p^2}{\omega^2}\right)} \tag{10.39}$$

and gives rise to an unattenuated wave which is part reflected and part transmitted. The bulk reflectivity of a metal can be evaluated numerically and is given by substituting Eq. (10.33) into Eq. (10.23). The result is shown in Fig. 10.2.

10.3.2 The Classical and Quantum Conductivity

One question the reader may ask at this stage is, how come it is possible to describe the optical properties of an electron gas with classical methods and get the right answers? The answer to this question is that if one carries through the fully quantum mechanical derivations of the above results, one arrives in the limit of weak scattering, to essentially the same answers. The quantum mechanical derivation does however tell us two new important things: (1) that the lifetime τ entering the Drude theory is not classical friction, but the quantum mechanical coherence time of electrons. It is the average time a particle stays in an eigenstate before it is scattered out of it by a phonon or an impurity potential, defect, etc. and (2) that true quantum effects become important when the electron gas is not treatable in the nearly free electron approximation anymore. If the metallic system is an alloy, or a liquid metal, or an amorphous medium, for example, then the quantum description matters very much. Indeed in this limit, the improved quantum mechanical theory tells us that there is a serious modification which has to be made to the Drude result. The necessary change is to replace the carrier density n_c with the expression:

$$n_c \to \frac{1}{2} g(E_f) m^* |v_f|^2$$
$$\sigma(0) = g(E_f) E_F q^2 \tau / m^* \tag{10.40}$$

The above new equation for the conductivity signifies that the carrier density in the Drude formula is in reality the density of states at the Fermi level times the Fermi energy (Fermi velocity squared times ½ effective mass). In the nearly free electron gas, the two are identical and the right-hand side of Eq. (10.40) is exactly n_c. But in a more complex metal, the density of states at the Fermi energy can be very different from the free electron form, both in its energy dependence and its value. Indeed if the density of states at the Fermi level is zero, or below a "minimum number," then the electron gas has no mobile carriers which can respond to a field, and the system does not conduct at all! In classical physics, electrons do not obey Fermi statistics, and all carriers can participate in conduction. Not so in quantum physics, Eq. (10.39) says that only the ones near the Fermi level can respond to a small electric field. Changing the density of states at the Fermi level therefore strongly affects the transport properties and consequently also the optical properties. This observation is particularly important for low-dimensional systems, where it is possible to engineer and externally manipulate the band structure and therefore the density of states at E_f.

The reader is referred to Madelung's (1978) and Ziman's (1964, 1969) books in the further reading section for a more detailed discussion of quantum transport.

10.4 The Bound and Valence Electron Contributions to the Permittivity

10.4.1 Time-Dependent Perturbation Theory

Consider now the influence of bound electrons on the optical properties. When bound charges are subject to an electric field, they will also be displaced, but not freely, and not to "infinity," as the frequency tends to zero. For bound electrons, the external field is only a small perturbation, which gives rise to polarization of the bonds and orbits, and we can apply methods of quantum mechanical perturbation theory. We consider therefore the effect of the time-dependent external field as an additional new term in the total energy or Hamiltonian of the system:

$$V(t) = -q \vec{E} . \vec{r} \tag{10.41}$$

The next step is to solve the time-dependent Schrödinger equation in Chap. 4 in the presence of this new term. Previously the Hamiltonian was time independent, and we could therefore write the unperturbed solutions in the usual way as shown in Chap. 4, namely, as the set:

$$\Psi_n\left(\vec{r},t\right) = \Phi_n\left(\vec{r}\right)e^{-iE_nt/\hbar} \tag{10.42}$$

with energy eigenvalues E_n. In the presence of the perturbation, the electrons are no longer in their stationary sates but can now admix with other, higher-lying excited states and change their orbital configurations and in principle also undergo transitions into these excited states. The change of spatial configuration is just what polarization is in the classical sense, and the transition into excited states is what we call absorption of energy from the light beam. We shall now see how polarization and absorption can be computed in quantum mechanics. We do this by assuming without loss of generality that the system was in its ground state g for $t < 0$, and then the effect of the perturbation applied at $t = 0$ is to generate a new electronic configuration which is a superposition of the ground state and all the other excited states of the system. The new wavefunction is a solution of the time-dependent Schrödinger equation in the presence of the coupling term described in Eq. (10.41). We emphasize that the principle of superposition is rigorously true and part of the principles of quantum mechanics we discussed in Chap. 4. So we can write for $t > 0$:

$$\Psi\left(\vec{r},t\right) = \Phi_g e^{-iE_gt/\hbar} + \sum_{n\neq g} c_n(t)\Phi_n e^{-iE_nt/\hbar} \tag{10.43}$$

where g denotes the ground state and n the excited states. The next step is to determine the new admixture coefficients $c_n(t)$. We do this by substituting Eq. (10.43) into the time-dependent Schrödinger equation (see Eq. 4.4a)). On one side we take the derivative with respect to time to obtain:

$$i\hbar\frac{\partial\Psi}{\partial t} = E_g\Phi_g e^{-iE_gt/\hbar} + \sum_{n\neq g} E_nc_n(t)\Phi_n e^{-E_nt/\hbar} + \sum_{n\neq g} i\hbar\frac{\partial c_n}{\partial t}\Phi_n e^{-iE_nt/\hbar} \tag{10.44}$$

On the other side of the Schrödinger equation, we have:

$$\{H_0 + V(t)\}\Psi\left(\vec{r},t\right) = E_g\Phi_g e^{-iE_gt/\hbar} + \sum_{n\neq g} E_nc_n(t)\Phi_n e^{-iE_nt/\hbar} - q$$
$$\vec{r}\cdot\vec{E}_0\left(e^{-i\omega t} + e^{-i\omega t}\right)\Psi\left(\vec{r},t\right) \tag{10.45}$$

We now equate Eqs. (10.44 and 10.45) and cancel the common terms. This leaves the last terms of the right-hand side of Eqs. (10.44 and 10.45) as equal to each other. Now we multiply the new equation on both sides with $\Phi_j^* e^{iE_jt/\hbar}$ and integrate over space. This operation eliminates all orthogonal terms, because we are using the fact that states belonging to different eigenvalues are orthogonal to each other (see Eq. 4.6). We also drop all terms which involve the product of the perturbation $V(t)$ and a coefficient $c_l(t)$ because such terms are necessarily of second order or above in the strength of the perturbation. The orthogonality rule, and the first-order

perturbation approximation, only leaves one term in the sum of the last term on the right-hand side of Eq. (10.45) which now gives:

$$i\hbar\frac{\partial c_j}{\partial t} = -\int d\ \vec{r}\ \Phi*_j(\vec{r})q\ \vec{r}\ \vec{E}_0\left(e^{i\omega t}+e^{-i\omega t}\right)e^{i\left(E_j-E_g\right)t/\hbar}\Phi_g(\vec{r}) \tag{10.46}$$

This can be integrated to give:

$$c_j(t) = -q\vec{E}_0\cdot\vec{r}_{jg}\left[\frac{1-e^{i\left(\hbar\omega+E_j-E_g\right)t/\hbar}}{\hbar\omega+\left(E_j-E_g\right)}-\frac{1-e^{i\left(-\hbar\omega+E_j-E_g\right)t/\hbar}}{\hbar\omega-\left(E_j-E_g\right)}\right] \tag{10.47}$$

where the position matrix element is:

$$\vec{r}_{jg} = \int d\ \vec{r}\ \Phi*_j(\vec{r})\ \vec{r}\ \Phi_g(\vec{r}) \tag{10.48}$$

For simplicity we assume that the wave is polarized in the x-direction so the first factor reduces to $qE_0^x x_{jg}$. Equation (10.48) is, apart from a factor q, the matrix element of the dipole moment of the electron; it is a measure of how much the excited state j has ground state g character mixed into it when acted on by the position coordinate. The matrix element of an operator Eq. (10.48), in this case the displacement, $\vec{r}_{\alpha\beta}$, is sometimes also written in the Dirac notation $\left\langle\alpha\left|\vec{r}\right|\beta\right\rangle$.

The above results now allow us to compute how the applied field polarizes the bound electron system. By definition the induced time-dependent dipole moment $P_x(t)$ is given by the charge q times the expectation value of the position operator:

$$P_x(t) = -q\int d\ \vec{r}\ \Psi^*(\vec{r},t)x\Psi(\vec{r},t) \tag{10.49}$$

Substitute the solution from the wavefunction, and keep only the linear terms in the coefficients which immediately give us:

$$P_x(t) = -\sum_j q\left(x_{gj}c_j(t)e^{-i\omega_j t}+x_{jg}c_j{}^*(t)e^{-\omega_j t}\right) \tag{10.50}$$

$$P_x(t) = \sum_j q^2\left|x_{gj}\right|^2\left(\frac{1}{E_{j0}-\hbar\omega}+\frac{1}{E_{j0}+\hbar\omega}\right)E_0^x\left(e^{-i\omega t}+e^{-i\omega t}\right) \tag{10.51}$$

From the dipole moment induced by the field, we can now deduce the polarizability in the usual way:

$$\alpha_p(\omega) = \sum_j q^2\left|x_{gj}\right|^2\frac{2E_{j0}}{E_{j0}^2-(\hbar\omega)^2} \tag{10.52}$$

and by introducing the oscillator strength F_j:

$$F_j = \frac{2m_0}{\hbar^2} E_{jg} \left| x_{gj} \right|^2 \tag{10.53}$$

We can rewrite the ground state polarizability in an elegant form:

$$\alpha_p(\omega) = \frac{q^2}{m_0} \sum_j \frac{F_j}{\omega_{jg}^2 - \omega^2} \tag{10.54}$$

with $\omega_{jg} = (E_j - E_g)/\hbar$. The significance of this expression becomes clear when we note that the oscillator strengths obey a simple sum rule:

$$\sum_j F_j = 1 \tag{10.55}$$

This sum rule is important. It is a check of consistency and follows from two quantum mechanical identities. The momentum position commutation relation:

$$xp_x - p_x x = i\hbar \tag{10.56}$$

and taking the expectation value of this equation and expanding over a complete set of intermediate states:

$$i\hbar = \sum_l \left(x_{il} p_{li,x} - p_{il,x} x_{li} \right) \tag{10.57}$$

and using an identity from Heisenberg equation of motion which reads:

$$p_{ij,x} = x_{ij} \left(E_j - E_i \right) m_0 / i\hbar \tag{10.58}$$

Substituting Eq. (10.58) into Eq. (10.57) gives the sum rule. Now that we know the bound electron polarizability, we can compute the relative permittivity by considering the polarizability of N_b of such atoms or molecules per unit volume.

$$\varepsilon(\omega) = 1 + \frac{N_b q^2}{\varepsilon_0 m_0} \sum_j \frac{F_j}{\omega_j^2 - \omega^2} \tag{10.59}$$

The sum now runs over the eigenstates of one such elementary unit, i.e., an atom or a molecule. In the zero frequency limit, we have:

$$\varepsilon(0) = 1 + \sum_j \frac{F_j \omega_p^2}{\omega_j^2} \tag{10.60}$$

And in the high-frequency limit, when the light energy exceeds all bound-to-bound transitions, we recover the corresponding Drude result:

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \tag{10.61}$$

which also implies that close to the plasma frequency, the permittivity can be negative, and the refractive index is purely imaginary implying from Eq. (10.23) perfect reflection.

10.4.2 Real Transitions and Absorption of Light

So far we have not considered what happens when the energy of the photon matches the energy difference between two bound levels. From Eq. (10.49), we should expect an infinite response. But what does this mean? When we have matching of energies, we should expect the electron to reach the excited state and the photon to be absorbed. In order to track such a transition mathematically, we go back to Eq. (10.47) and evaluate the probability that the particle is in the excited state j at time t having started at $t = 0$ in the ground state. From Eq. (10.47) we note that in the expression for $c_j(t)$, there are two terms, one corresponding to the possibility of absorption, namely, a resonance when $\hbar\omega = \hbar\omega_j$, and one corresponding to emission. For simplicity we keep the absorption term only so we have:

$$\left|c_j(t)\right|^2 \sim \left|\frac{qx_{gj}}{\hbar}E_0^x\right|^2 \frac{\sin^2\left(\omega_j - \omega\right)t/2}{\left(\omega_j - \omega\right)^2} \tag{10.62}$$

The right-hand term or *sine* function is strongly peaked at $\omega_j = \omega$ and decays strongly with frequency; it is a well-known function of mathematical physics and is best analyzed if, instead of the probability, we consider the probability per unit time of finding the particle in the excited state j that is divided by time t to study $W_{gj} = |c_j(t)|^2/t$. Dividing the right-hand side of Eq. (10.62) by t, and letting time go to infinity, gives us a function which we recognize to be the well-known Dirac delta function:

$$t \to \infty \Rightarrow \frac{\sin^2\left[\left(\hbar\omega_j - \hbar\omega\right)t/2\hbar\right]}{t\,\hbar^2\left(\omega_j - \omega\right)^2/4} = \frac{2\pi}{\hbar}\delta\left(\hbar\omega_j - \hbar\omega\right) \tag{10.63}$$

The Dirac delta function $\delta(x)$ has the property that:

$$\int_{-\infty}^{\infty} dx\delta(x) = 1 \tag{10.64}$$

And also as the imaginary part of the fraction:

$$\mathrm{Im}\left(\frac{1}{x-i\eta}\right)=\pi\delta(x) \tag{10.65}$$

with infinitesimal η. So basically Eq. (10.62) contains the statement that the particle can end up in an excited state if energy is conserved in the long time limit. Although the Heisenberg uncertainty relation allows energy not to be conserved at short times, to complete the transition, and to make a temporary admixture real, energy conservation must be satisfied in the long time limit.

We can summarize this result in the form known as the Fermi's golden rule which states that if a particle is subject to perturbation of the form $2V(r)\cos\omega t$, then the probability per unit time of finding it in an eigenstate j given that it started in g at $t=0$ is given by the formula:

$$W_{gj}=\frac{2\pi}{\hbar}\left|\int d\,\vec{r}\;\Phi_j{}^{*}V\left(\vec{r}\right)\Phi_g\right|^2\delta\left(\hbar\omega-E_j+E_g\right) \tag{10.66}$$

Now we can understand the meaning of the resonances in the permittivity expression Eq. (10.59). They do indeed indicate absorption processes, and the way to take care of the singularity is to introduce the notion of a lifetime. Clearly when excited, the electron can recombine back down again so it has a finite lifetime in the excited state, and by Heisenberg uncertainty principle, because of this time uncertainty, it has a finite energy uncertainty or energy broadening. There is a broadening associated with each level j, and the lifetime is measured in Hz. The broadening introduces a complex number in the denominators of Eq. (10.47) so that the relative permittivity becomes the complex function ($T=0$ K):

$$\varepsilon_b(\omega)=1+\frac{N_bq^2}{\varepsilon_0m_0}\sum_j\frac{F_j}{\omega_j^2-\omega^2-i\omega\Gamma_j} \tag{10.67}$$

This function has a real and an imaginary part. The imaginary part, we know, is related to the absorption coefficient, and this time it is not the joule heating of free electrons as in Drude theory but the absorption of photons by bound electrons in the solid. We are now in the position to write down an expression for the relative permittivity of the solid including both bound N_b and free electrons n_c:

$$\varepsilon(\omega)=1+\frac{N_bq^2}{\varepsilon_0m_0}\sum_j\frac{F_j}{\omega_j^2-\omega^2-i\omega\Gamma_j}-\frac{n_cq^2}{\varepsilon_om^*}\left(\frac{\omega\tau-i}{\omega^2\tau^2+1}\right) \tag{10.68}$$

At this stage it is also useful to generalize the bound relative permittivity to finite temperatures, allowing the light to admix bound levels up and admix thermally excited levels down in energy, to find (Γ_{ij} largest of the two widths and f_l is the Fermi-Dirac function):

$$\varepsilon_b(\omega)=1+\frac{N_bq^2}{\hbar^2\varepsilon_0}\sum_{i\neq j}\frac{\hbar\left|x_{ij}\right|^2\left(f_i-f_j\right)\left(\omega_j-\omega_i\right)}{\left(\omega_j-\omega_i\right)^2-\omega^2-i\omega\Gamma_{ij}} \tag{10.69}$$

10.4.3 The Permittivity of a Semiconductor

We can apply these results to a semiconductor. Consider a direct bandgap semiconductor with no free carriers for the sake of simplicity. In this case the bound electrons are in the valence band, and the quantum label j becomes a Bloch $\vec{k}$-state, and the number of orbital N_b/volume falls under the Bloch integral $\vec{k}$. The transitions that the light can induce are from valence to conduction band and involve a negligible momentum of the light wave. For band-edge absorption, this is only possible with direct bandgap materials (see Fig. 5.17). The indirect bandgap systems will be discussed later on in this chapter. In direct bandgap materials, or for sufficiently high photon energy, Eq. (10.67) means that the permittivity involves to a good approximation only the vertical $\vec{k}$-valence to same $\vec{k}$-conduction band admixtures. We also assume that the valence band is full and the conduction band is empty so that we have ($T = 0$ K):

$$\varepsilon_s(\omega) \sim 1 + \frac{q^2}{\varepsilon_0 m_0} \sum_{\vec{k}} F_{\vec{k}} \frac{1}{\left(\omega_{\vec{k},c} - \omega_{\vec{k},v}\right)^2 - \omega^2} \tag{10.70}$$

where the Bloch sum over the occupied states is normalized by the volume and defined as:

$$\sum_{\substack{\vec{k} \\ v}} = N_b \tag{10.71}$$

with N_b denoting the effective number of bound eigenstates per unit volume. At $\omega = 0$, the largest contributions in this sum are from the band-edge states, so the denominator can be replaced by the bandgap $E_g/\hbar$, and the oscillator strength for the vertical band-to-band transition $F_{\vec{k}}$ is to a good approximation reducible under the sum to give the total valence band electron density and therefore the expression:

$$\frac{q^2}{m_0 \varepsilon_0} \sum_{k} F_{\vec{k}} \sim \frac{N_b q^2}{\varepsilon_0 m_0} = \left(\omega_p^b\right)^2 \tag{10.72}$$

$$\varepsilon_s(0) \sim 1 + \left(\frac{\hbar \omega_p^b}{E_g}\right)^2 \tag{10.73}$$

where ω_p^b is the effective bound electron plasma frequency and can be obtained by comparison with the experiment. It should be roughly a factor $\frac{E_g}{E_{B,v}}$ ($E_{B,v}$ is the valence band width) smaller than the absolute valence band plasma frequency. This expression is valid for the low-frequency permittivity of a semiconductor of energy gap E_g. Given that a bandgap can typically be $\sim 3 \times 10^{14}$ Hz, we see that the low-frequency limit can go a long way. So in the range 0~10^{11} Hz, for example, the zero frequency

form is quite adequate, and for a doped semiconductor, the bound valence band contribution can be combined with the free electron contribution.

At finite temperature, the above expression is still a good approximation in a wider gap semiconductor, but the full generalization for finite temperature, substituting for the oscillator strength, and including the broadening, is in fact:

$$\varepsilon_s(\omega) \sim 1 + \frac{q^2}{\hbar\varepsilon_0} \sum_{\vec{k}} \left| x_{\vec{k}c,\vec{k}v} \right|^2 \frac{\left(\omega - \omega_{\vec{k}c} + \omega_{\vec{k}v}\right) - i\gamma}{\left(\omega - \omega_{\vec{k},c} + \omega_{\vec{k},v}\right)^2 + \gamma^2} \left[f\left(E_{\vec{k}v}\right) - f\left(E_{\vec{k}c}\right) \right] \tag{10.74}$$

where the sum is now over the $\vec{k}$ index normalized per unit volume. The x-position matrix element has to be evaluated using the valence and conduction band Bloch functions. Fortunately and to a good approximation, this matrix element can be calculated using Kane theory to give us the result (Rosencher and Vinter 2002):

$$|x_{kvkc}|^2 = \left| \int d\,\vec{r}\; \Psi *_c\left(\vec{k}\right) x \Psi_v\left(\vec{k}\right) \right|^2 = \frac{1}{3} \frac{\hbar^2}{E_g^2} \frac{E_P}{m_0} \tag{10.75}$$

where E_P is the Kane parameter and a number which varies only slightly between 20 and 25 eV in most semiconductors (see also Sect. 5.7). This powerful last equation now allows us to compute the permittivity for most situations of interest in semiconductor physics. All we need for Eq. (10.74) is the density of band states which as we know is usually well described in the nearly free electron approximation.

10.4.4 The Effect of Bound Electrons on the Low-Frequency Optical Properties

We have seen that bound electrons usually contribute frequency dependence to the permittivity only at high frequencies. When we consider both free and bound carriers, we must go back and see how one affects the other. One of the important consequences of ε_b on the free carrier response is in the regime $\omega\tau \gg 1$ discussed previously for free carriers only. The combined permittivity in this regime is approximately real, but the bound electron contribution is significant, so that the refractive index now becomes:

$$\bar{n}(\omega) = \left((1 + \varepsilon_b) \left(1 - \frac{\omega_p^2}{\omega^2 (1 + \varepsilon_b)} \right) \right)^{1/2} \tag{10.76}$$

or as is the notation of some other authors, one can also replace:

$$\varepsilon(\infty) = 1 + \varepsilon_b \tag{10.77}$$

One can think of Eq. (10.76) as a renormalization of the plasma frequency of the electrons to $\omega_p^2 \rightarrow \omega_p^2/(1+\varepsilon_b)$. This is a real effect because the electrons are now oscillating in a medium in which the electric field of the restoring force is screened by the permittivity of the bound carriers. The low-frequency permittivities of some important semiconductors are given in Appendix 4, for example, GaAs, $\varepsilon_b = 13.1$; Si, $\varepsilon_b = 11.9$; and C, $\varepsilon_b = 5.7$. From Eq. (10.73), it follows that the large bandgap materials are expected to have the lower permittivity, and this is in general observed.

10.5 The Optical Absorption in Semiconductors

10.5.1 Absorption Coefficient

The optical absorption of a direct bandgap semiconductor is given by the imaginary part of the permittivity Eq. (10.67) (Fig. 10.3).

This is a sum of energy-conserving transitions described by matrix elements which take an electron from the valence band vertically up (i.e., same $\vec{k}$ value) to the conduction band. The number of such terms is therefore directly proportional to the number of available band states. Thus the optical absorption properties of semiconductors are intimately related to the density of allowed states in the conduction and valence bands.

The absorption process is characterized by the absorption coefficient, $\alpha(\omega)$, which is usually expressed in units of cm^{-1} or m^{-1} in MKS, as used in this book. This quantity depends on the incident photon energy $\hbar\omega$ and expresses the ratio of the number of photons actually absorbed by the crystal per unit volume per second, to the number of incident photons per unit area per second. The calculation of the absorption coefficient for a direct bandgap material resembles that of the density of states but takes into account the $E-\vec{k}$ relationships in both the conduction band for electrons (with effective mass m_e) and the valence band for holes (with effective mass m_h). This consideration results from two important conservation laws that rule

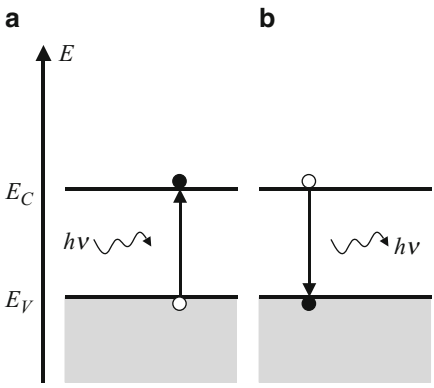


Fig. 10.3 Electronic transition, (**a**) from the valence band to the conduction band resulting from the absorption of a photon, (**b**) from the conduction band to the valence band resulting into the emission of a photon

the optical absorption process: (i) the total energy (electron+hole+photon) must be conserved and (ii) the total momentum or wavevector must also be conserved. Assuming that in Eq. (10.70), the oscillator strength $F_{\vec{k}}$ is only a weak function of $\vec{k}$ which allows us to take the imaginary part of the permittivity as the delta function sum to obtain for absorption, with:

$$\mathrm{Im}\left(\frac{1}{\hbar\omega - E_C\left(\vec{k}\right) + E_V\left(\vec{k}\right) - i\eta}\right) = \pi\delta\left[\hbar\omega - E_C\left(\vec{k}\right) + E_V\left(\vec{k}\right)\right] \tag{10.78}$$

and therefore having split up the expression in Eq. (10.70), we have:

$$2\bar{n}\kappa = \frac{\hbar^2 q^2}{m_0\varepsilon_0} F_{vc} \int \frac{1}{2\hbar\omega} d\vec{k}\ \delta\left(\hbar\omega - E_C\left(\vec{k}\right) + E_V\left(\vec{k}\right)\right) \tag{10.79}$$

The delta function sum is called the joint density of states per volume and can be evaluated as the ordinary density of states by introducing the reduced mass via (remember the valence band energy is defined negative):

$$\hbar\omega = \frac{\hbar^2 k^2}{2m_e} + \frac{\hbar^2 k^2}{2m_h} = \frac{\hbar^2 k^2}{2}\left(\frac{1}{m_r}\right) \tag{10.80}$$

The absorption coefficient can then be found to be proportional to the density of states, with the effective mass m^* replaced by the reduced effective mass defined as:

$$m_r^* = \frac{m_e m_h}{m_e + m_h} \tag{10.81}$$

For example, in a three-dimensional bulk semiconductor structure with direct bandgap:

$$2\bar{n}\kappa = \frac{\hbar q^2}{2\omega m_0\varepsilon_0} F_{vc}\left\{\frac{1}{2\pi^2}\left(\frac{2m_r^*}{\hbar^2}\right)^{3/2}\sqrt{\hbar\omega - E_g}\right\} \tag{10.82}$$

where by definition, the absorption coefficient is $\alpha = \frac{\omega}{\bar{n}c} 2\bar{n}\kappa$ and where $\hbar\omega$ is the incident photon energy, E_g is the energy gap of the semiconductor, and F_{vc} can be evaluated using Kane theory Eq. (10.75) (see also Sect. 5.6).

A word of caution: when using approximation methods such as Kane theory, it can happen that the oscillator strength defined using the bare mass as in Eq. (10.53) exceeds 1, which is inconsistent with the sum rule. This is because the sum rule should really be evaluated within the same scheme so that m_0 in Eq. (10.53) should be replaced by the Kane m^* (see Sect. 5.7). The expression in the curly bracket on the right-hand side of Eq. (10.82) is called the electron-hole or joint density of states because it takes into account the density of states in both the conduction and valence bands.

In reality, the absorption spectra do not reproduce exactly the joint density of states because there are other processes which contribute to absorption as well. These are due to photons coupling to lattice vibrations, i.e., electron-phonon interactions and also excitonic effects. Let us first consider the excitonic contribution.

10.5.2 Excitonic Effects

Let us now consider excitonic effects. An electron excited into the conduction band is a negatively charged particle in a neutral medium which will interact with the resulting hole created in the valence band (positively charged particle). In other words when light creates an e-h pair, it is not yet a free pair. This pair of charged particles is created locally, and they attract each other by the Coulomb force. They form a unit called the exciton. In an exciton, the electron and the hole attract each other and move together as a single particle consisting of a coupled (i.e., not free) electron-hole pair. This pair resembles a hydrogen atom where the role of the nucleus is played by the hole.

An exciton has two degrees of freedom: the relative motion of the electron and the hole and the motion of the exciton as a single unit. As in the case of the hydrogen atom, the relative motion is quantized, and the energy spectrum of an exciton consists of discrete energy levels in the bandgap corresponding to the ground state and the excited states of an exciton. But unlike in hydrogen, the pair is moving in a medium which has a finite polarizability as we have just seen above. So the Coulomb potential is screened by the medium. Using the results of section 0 for the hydrogen atom, we obtain the energy associated with the relative motion of an exciton:

$$E_n = -\frac{E_{Ry}}{n^2} \tag{10.83}$$

where $n = 1,2,\ldots$ is an integer and E_{Ry} is the exciton Rydberg energy. This shows that, similarly to the hydrogen atom, the energy spectrum of the relative motion of an exciton consists of discrete levels. Each level is indexed by a main quantum number n, and the wavefunctions are characterized by orbital quantum numbers $l = 0, 1, \ldots, n-1$ and magnetic quantum numbers $m = -l, -l+1, \ldots, l$. The Rydberg energy is given by:

$$E_{Ry} = \frac{m_r^* q^4}{8(\varepsilon_r \varepsilon_0 h)^2} \tag{10.84}$$

with ε_r is the real part of the zero-frequency relative permittivity or the dielectric constant of the material and ε_0 is the permittivity of free space and h is Planck's constant. Furthermore, by defining an exciton Bohr radius, a_B, derived from Eq. (1.3) such that:

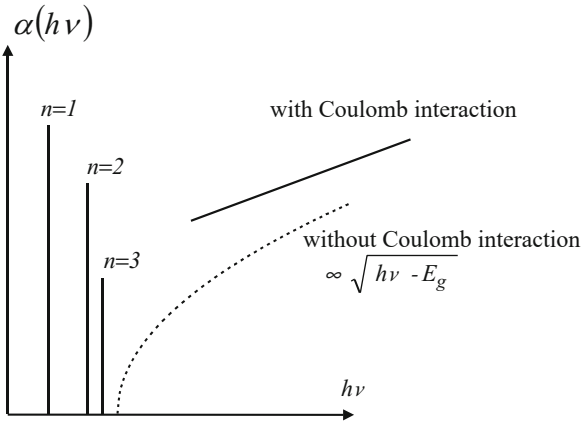


Fig. 10.4 Excitonic absorption peaks ($n = 1,2,3$) in the optical absorption spectra of a bulk semiconductor (3D). These peaks are located inside the energy gap. In addition, the effect of coulombic interaction between electrons and holes on the absorption coefficient is shown

$$a_B = \frac{\varepsilon_r \varepsilon_0 h^2}{\pi m_r^* q^2} \tag{10.85}$$

we can rewrite the Rydberg energy as:

$$E_{Ry} = \frac{q^2}{8\pi\varepsilon_r\varepsilon_0 a_B} = \frac{\hbar^2}{2m_r^* a_B{}^2} \tag{10.86}$$

The first fraction is similar to Eq. (1.5) and expresses the hydrogen atom analogy for the exciton. The energy spacing between the ground state exciton level and the bottom of the conduction band is called the exciton binding energy and physically represents the energy needed to separate the electron and the hole into two free particles. We note that because of the permittivity of the host $\varepsilon_r = \varepsilon_b(0) > 1$, the binding energy is considerably reduced compared to the hydrogen atom. Given that for a semiconductor like silicon, $\varepsilon_s \sim 10$, and this is true for most semiconductors of interest ($\varepsilon_b \sim 10 - 15$), we have a reduction of energy of ~100–300 from 13 to ~0.13 eV and less.

Excitons can be efficient absorbers of light. When excitons are involved in the optical absorption process, the absorption spectrum exhibits additional sharp peaks within the energy gap, near the bandgap energy (E_g), corresponding to the excitonic energy levels. This is illustrated in Fig. 10.4 for a bulk semiconductor (3D). In addition, even at higher energies, deep inside the conduction band where excitons are typically not encountered, the absorption coefficient is still influenced by the Coulomb interaction between electrons and holes.

It should be noted that in bulk semiconductors, the presence of excitons has been verified only at cryogenic temperatures. This is because an exciton has a small binding energy, and electron-phonon interactions can, at higher temperatures, easily break up the exciton into free electrons and holes, i.e., the lifetime of an exciton is very short at high temperatures.

However, in low-dimensional structures, one can observe excitonic effects at much higher temperatures because the spatial confinement reduces the screening

efficiency and enhances the binding of the pair; they have a smaller chance to escape and thus a larger exciton binding energy. We shall see this in Chap. 12.

10.5.3 Direct and Indirect Bandgap Absorption

The formalism for the optical permittivity of semiconductors above applies mainly for the direct bandgap materials because it assumes transitions with zero momentum exchange. This includes the important class of materials such as GaAs and InAs. Now let us consider the indirect bandgap systems.

In the chapter where we discussed the band structure of semiconductors, recall Fig. 5.17 in Chap. 5; we encountered two distinct classes of materials: the direct and indirect bandgap materials**.** Semiconductors like Si and Ge have indirect bandgaps. That means that the lowest photon energy that can be absorbed necessarily involves a change of momentum, and this process is not included in the formalism of Eq. (10.74).

From Fig. 5.17 for Ge, we see that the lowest-energy absorption is one where an electron is taken out of the top of the valence band at the Γ point and put into the lowest energy in the conduction band at the X point. The momentum change is substantial and cannot be supplied by the photon; it must come from another source. The most obvious one is the phonon bath. Phonons can couple to the photons and make the transition happen. They can do this in absorption or in emission of a phonon. Energy and momentum can be satisfied in particular with optical phonons where the energy dispersion with momentum is weak and can be neglected for most purposes. Energy conservation gives:

$$E_c\left(\vec{k}+\vec{Q}\right)=E_v\left(\vec{k}\right)\pm\hbar\Omega+\hbar\omega \tag{10.87}$$

where the required momentum $\vec{Q}$ is fixed by the band structure**.** The process can be one of the emissions in which case the photon needs more energy than the indirect bandgap. The emission process is weakly dependent on temperature and involves the factor $1/\left(e^{\hbar\Omega/k_bT}-1\right)+1=N(\omega)+1$. Phonon absorption, on the other hand, can happen with photon energies less than the indirect bandgap, but only if such phonons are excited, so here we have a Bose factor $N(\omega)$ which is temperature dependent. In summary after doing the integrations in the corresponding Fermi's golden rule formulae, one arrives at the two indirect absorption coefficients which have the form (assisted with the emission and absorption of an optical phonon, respectively):

$$\begin{aligned}\alpha_{ep}&=A_e\frac{\left(\hbar\omega-E_{ind}^g-\hbar\Omega\right)^2}{1-e^{-\hbar\Omega/k_bT}}\theta\left(\hbar\omega-E_{ind}^g-\hbar\Omega\right)\\ \alpha_{ep}&=A_a\frac{\left(\hbar\omega-E_{ind}^g+\hbar\Omega\right)^2}{-1+e^{\hbar\Omega/k_bT}}\theta\left(\hbar\omega-E_{ind}^g+\hbar\Omega\right)\end{aligned} \tag{10.88}$$

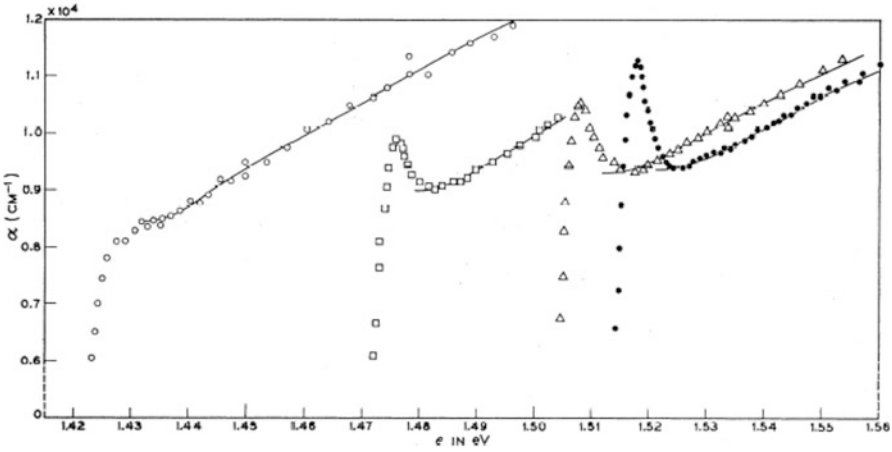


Fig. 10.5 Band-edge absorption of GaAs showing also the evolution of the exciton absorption for different temperatures left to right: 294 K;186 K; 90 K; 21 K (Reprinted figure with permission from Sturge 1962, p. 768, Fig. 3. Copyright 1962 by the American Physical Society)

The $A^{'}s$ are constants, and the theta function θ is zero when the argument is less than zero and one otherwise (Peyghambarian et al. 1993). Note the different (squared) behaviors of the band-edge absorption Eq. (10.88) with photon energy when compared with the direct bandgap case Eq. (10.82) (square root).

Figures 10.5 and 10.6 illustrate the absorption edges of GaAs and Ge. The GaAs data is plotted on a linear scale and the Ge data logarithmically so that one can see the crossover from indirect to direct absorption at the inflection point of the curve.

When a phonon is needed, the transition is more complex, involves three bodies, and is therefore also less efficient. When an electron is excited in the conduction band with high energy, so that the direct $\vec{k}=\vec{0}$ transition is possible, it will in general thermalize down very quickly to the indirect band edge, and light emission will only take place at the final recombination step at the lowest bandgap. In an indirect bandgap system, a phonon is needed, and therefore materials such as Ge and Si will be poor light-emitting systems (Fig. 10.6).

10.6 The Effect of Phonons on the Permittivity

10.6.1 Photon Polar Mode Coupling

We have so far included free and bound electrons contributions in the permittivity. We have discussed the effect of excitons, so now we must ask: what other processes can affect the optical response of a solid? Clearly at finite temperatures, the lattice atoms are thermally excited and vibrate. We have seen that the atomic bonds can be polar, and the lattice dipoles can vibrate and be stimulated to vibrate by light waves. This means that in particular, it is also possible for such polar lattice vibrations to

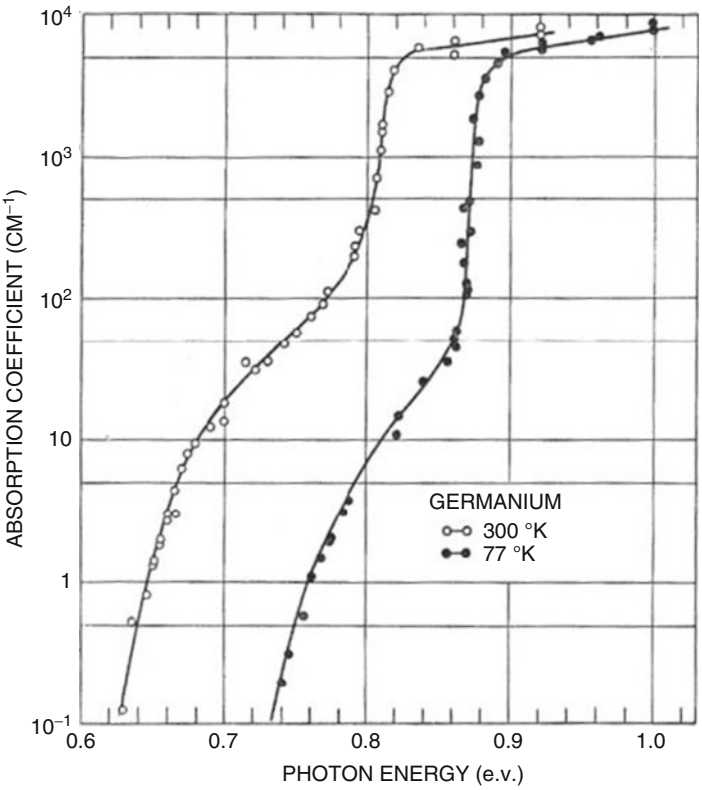


Fig. 10.6 The band-edge absorption of Ge on a logarithmic scale. Note the change of behavior at 102 cm^{-1} from indirect to direct band-to-band transitions (Reprinted from Newman and Tyler (1959), p. 58, Fig. 1. Copyright (1959), with permission from Elsevier)

absorb energy from the light passing through the medium. The effect of light coupling to atomic motion is not negligible in semiconductors with polar modes and needs special treatment. The general treatment of photon-phonon coupling, i.e., including acoustic coupling and many phonon effects, is beyond the scope of this textbook. In this chapter, we will develop the methodology for the strongest interaction, namely, for the polar lattice.

To investigate the influence of atomic vibrations on the permittivity, we consider the two-atom model of lattice vibrations from Chap. 6. If the bond is polar, then the atoms in the bond carry a net charge and couple to the light wave. Furthermore the vibrating atoms or charge can reemit light and also give up its extra energy to other phonon modes. So we also introduce a damping term γ to take care of this effect. The equation of motion Eq. (6.5) now becomes:

$$\begin{aligned} M_1\frac{d^2u_n}{dt^2}+M_1\gamma\frac{du_n}{dt}-C(v_{n+1}+v_{n-1}-2u_n)&=-qE_0e^{-i\omega t}\\ M_2\frac{d^2v_n}{dt^2}+M_2\gamma\frac{dv_n}{dt}-C(u_{n+1}+u_{n-1}-2v_n)&=qE_0e^{-i\omega t}\end{aligned} \tag{10.89}$$

where we have assumed that the M_1 mass is negatively charged and M_2 positively charged. The damping term is here, as before, proportional to the velocity.

We are not interested in the complete solution of this problem, so we focus only on those modes which could result in absorption or strong scattering of light, and we know that this is only possible when momentum is conserved. Since the photon only has negligible momentum to exchange, light can only excite or absorb phonons with a small momentum. It can absorb or emit acoustic and optical modes with small momentum exchange. With optical modes it is possible to excite relatively high-energy phonons with almost zero momentum. Indeed energy exchange can take place with optical modes near $k = 0$. So we focus only on those solutions to Eq. (10.89), namely, the ones at or near $k = 0$. The $k = 0$ optical phonon modes are the ones where the two sublattices move in phase relative to each other. We try a $k = 0$ mode:

$$\begin{aligned} u_n(t)&=A_1e^{-i\omega t}\\ v_n(t)&=A_2e^{-i\omega t}\end{aligned} \tag{10.90}$$

and find from Eq. (10.89):

$$\begin{aligned} [2C-M_1(\omega^2+i\gamma\omega)]A_1-2CA_2&=-qE_0\\ [2C-M_2(\omega^2+i\gamma\omega)]A_2-2CA_2&=-qE_0\end{aligned} \tag{10.91}$$

The solution is:

$$\begin{aligned} A_1&=\frac{-qE_0}{M_1\left[\Omega_+^2-(\omega^2+i\gamma\omega)\right]}\\ A_2&=\frac{qE_0}{M_2\left[\Omega_+^2-(\omega^2+i\gamma\omega)\right]}\end{aligned} \tag{10.92}$$

$$\Omega_+^2(k=0)=\frac{2C(M_1+M_2)}{M_1M_2} \tag{10.93}$$

Using this result we can now go back and compute the polarization induced by the light wave. Given N_l ion pairs per unit volume, we have the volume dipole moment:

$$P_l=-qN_l(u_n-v_n) \tag{10.94}$$

or:

$$P_l = qN_l(A_2 - A_1)e^{-i\omega t} \tag{10.95}$$

which then using Eq. (10.92) reduces in the limit of the pure ionic permittivity to:

$$\varepsilon_l(\omega) = 1 + \frac{qN_l}{E_0\varepsilon_0}(A_2(\omega) - A_1(\omega)) \tag{10.96}$$

$$\varepsilon_l(\omega) = 1 + \frac{q^2N_l}{\varepsilon_0 M_r}\frac{1}{\Omega_+^2 - (\omega^2 + i\gamma\omega)} \tag{10.97}$$

$$M_r = \frac{M_1M_2}{M_1 + M_2} \tag{10.98}$$

The optical phonon contribution to the permittivity has a real and imaginary part, from which one can evaluate the effect of optical phonons on light dispersion and absorption. We are now at last in a position to write down all the important contributions to the relative permittivity of a doped polar semiconductor as:

$$\varepsilon(\omega) = 1 + \left(\varepsilon_f(\omega) - 1\right) + (\varepsilon_b(\omega) - 1) + (\varepsilon_l(\omega) - 1) \tag{10.99}$$

where the polarizability contributions are added to give Eqs. (10.34, 10.67, and 10.97) and where it is understood that in a semiconductor, the bound contribution is the same as the formula in Eq. (10.69). With this theory we now can handle most situations of interest in semiconductor physics.

10.6.2 Application to Ionic Insulators

In this limit we neglect the free electrons, and it is again convenient to lump together all other lattice contributions into an$\varepsilon(\infty)$term and write the ion permittivity as being due to the transverse optical active mode denoted with frequency Ω_T:

$$\begin{aligned}\bar{n}^2 - \kappa^2 &= \varepsilon(\infty) + \frac{q^2N_l}{\varepsilon_0 M_r}\frac{\Omega_T^2 - \omega^2}{\left(\Omega_T^2 - \omega^2\right)^2 + \omega^2\gamma^2} \\ 2\bar{n}\kappa &= \frac{q^2N_l}{\varepsilon_0 M_r}\frac{\gamma\omega}{\left(\Omega_T^2 - \omega^2\right)^2 + \omega^2\gamma^2}\end{aligned} \tag{10.100}$$

Figure 10.7 shows the reflectivity R of an ionic insulator. The effect of the resonance on the reflectivity is to produce a sharp crossover from high to low reflectance as the photon energy is changed.

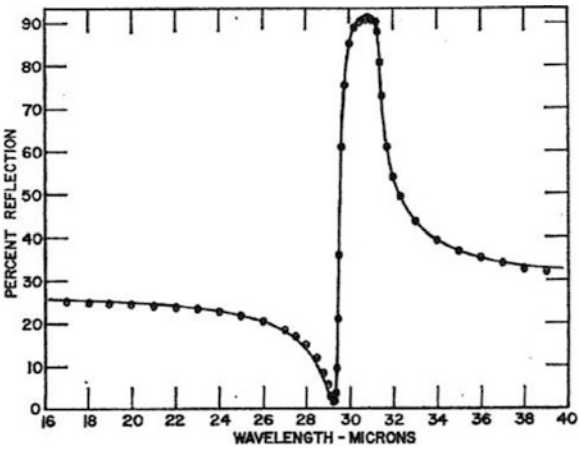


Fig. 10.7 Lattice reflection spectrum of AlSb. Points are experimental data; line is fit using the single oscillator model (Reprinted figure with permission from Turner and Reese (1962), p. 126, Fig. 4. Copyright 1962 by the American Physical Society)

10.6.3 The Phonon-Polariton

The real part of the refractive index due to the coupling with ions has a strong frequency dependence as can be seen in the previous figure and strongly modulates photons with frequencies in the neighborhood of the optical modes. Indeed the photon dispersion relation relating photon frequency and momentum k is:

$$\omega^2(k) = \left[\frac{ck}{\bar{n}(\omega)}\right]^2 \tag{10.101}$$

where $\bar{n}(\omega)$is given by the pair of Eq. 10.100. One can see that the refractive index changes with frequency so that the allowed frequencies of propagation of photons in the medium are solution of this equation, which can have several branches. Let us assume the damping is weak so that $\kappa(\omega) = 0$ in Eq. 10.100, and one has $\bar{n}(\omega)$ only so Eq. 10.101 becomes:

$$\omega^2(k)\left\{\varepsilon(\infty) + \frac{q^2 N_l}{\varepsilon_0 M_r}\frac{1}{\left(\Omega_T^2 - \omega^2(k)\right)}\right\} = c^2k^2 \tag{10.102}$$

which is a quadratic equation in ω^2 with two branches.

The frequency versus momentum of the physical roots is shown in Fig. 10.8 where $\Omega_L = \Omega_T\sqrt{\frac{\varepsilon(0)}{\varepsilon(\infty)}}$ turns out to be the longitudinal phonon frequency, and the zero frequency limit $\varepsilon(0)$ includes the zero frequency limit of the lattice term.

The excitation can be understood to be part photon and part phonon in its structure. Near $k = 0$ and at low frequency, it is mainly photon-like and basically follows the photon dispersion curve slowed down by the mainly bound electron refractive index $\sqrt{\varepsilon(\infty)}$ of course. Then, when the light energy reaches the optical mode energy of the phonon, a strong mixture of the two excitations takes place.

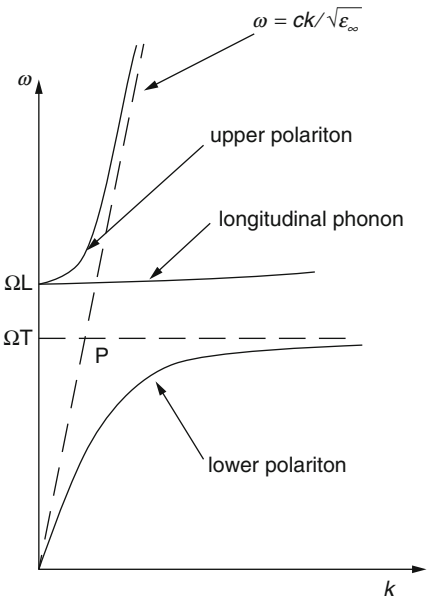


Fig. 10.8 The dispersion curve for a phonon-polariton (Peyghambarian et al. 1993, p. 98, Fig. 4.11. Modified with permission)

Here, the photon becomes a mixed state, part phonon and part photon; it gets slowed down in the process because the phonon is slow and almost localized. The group velocity of this combined particle can be much slower than light as one can see from the dispersion curve. At higher frequencies the two states demit because their energies no longer match, and the excitation acquires its photonic character again. This happens as we go up the k-axis and up in frequency. This photon which is crossing into a phonon-like mode is called a phonon-polariton. It is of great conceptual importance, as it allows regions of energy where photons can propagate at a much lower speed.

Photon-phonon coupling has many other very subtle aspects which we have not covered in this chapter. The reader is referred to the book by Seeger (1997) for a more specialized treatment. For example, whereas in III–V compounds, one does have polar bonds, the same is not true of other important classes of semiconductors such as silicon and germanium. Here phonon-phonon coupling and absorption are more subtle and involve higher-order processes. Whereas single-phonon excitations are forbidden by symmetry, higher order processed involving two and more phonons are allowed and give rise to rich absorption spectra.

10.7 Free Electrons in Static Electric Fields: The Franz-Keldysh Effect

So far we have assumed that the system in question is itself not subject to a strong electric or magnetic field. In this and the next sections, we consider the effect of an electric and magnetic field on the optical properties. Much of modern technology is

devoted to making optical systems for communication, displays, wavelength transformation, and computing. Optoelectronics is a very lively and exciting field and has acquired even more importance with the advent of nanotechnology. The basic element of all optical technology is the optical switch or optical transistor. How can one make a medium change its transparency or absorption properties by a simple low power electronic or magnetic switch? In order to understand how to design such a system using the right material, engineers need to understand what external fields do to the electronic structure of materials, and in particular they need to know how the optical properties of semiconductors behave when subjected to external fields.

Consider therefore a band of nearly free electrons in an electric field. We assume that we can use the effective mass approximation. When we previously considered the action of the electric field, it was in the context of electrical conduction, and it was good enough to treat the problem using a semiclassical approach. This is because the electric fields were small, and the dipoles generated were calculated to first order in field. Now we are looking at the effect of light on systems subject to strong electric fields, and we ask: what is new and important about strong electric fields? To answer this question, we first note that the external field is no longer a small perturbation on the wavefunctions. So we cannot use Drude-type theories but need to go back and solve the time-independent Schrödinger equation in the presence of an external electric field E_0^z applied in the z-direction, for example. It is understood that the motion in the x- and y-directions is nearly free electron-like, so that the total wavefunction and energy of the charge are separable:

$$\Psi_E(x,y,z) = e^{ik_x x} e^{ik_y y} \Phi(z) \tag{10.103}$$

$$E = \frac{\hbar^2}{2m^*}\left(k_x^2 + k_y^2\right) + E_z \tag{10.104}$$

The Schrödinger equation in the field direction becomes:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2 \Phi(z)}{\partial z^2} - qE_0^z z\Phi(z) = E_z \Phi(z) \tag{10.105}$$

Note that in the current formalism, the electric field is denoted E_0^z, while the energy associated with the wavefunction is denoted E_z. The wavefunctions which are solutions to these equations are called the Airy functions $Ai_v(z)$ with energy E_v and given by an integral representation:

$$Ai_v(z_v) = \frac{1}{\pi}\int_0^\infty \cos\left(\frac{s^3}{3} + sz_v\right) ds \tag{10.106}$$

$$z_v = \left(\frac{2m^* qE_0^z}{\hbar^2}\right)^{1/3}\left(z - \frac{E_v}{qE_0^z}\right) \tag{10.107}$$

The normalized eigenstates of Eq. (10.105) labeled with their energies are:

$$\Phi_{E_v} = \left(\frac{2m^*}{\hbar^2}\right)^{1/3}\left(\frac{1}{qE_0^z}\right)^{1/6} Ai\left[\left(\frac{2m^* qE_0^z}{\hbar^2}\right)^{1/3}\left(z - \frac{E_v}{qE_0^z}\right)\right] \tag{10.108}$$

The eigenfunctions can be thought of as starting at each lattice site, one for each site, at a distance a along the z-axis, so that $E_v/qE_0^z = av$ where v is an integer in the range $\{\infty, -\infty\}$, and av defines the origin of the v_{th} Airy state. The Airy function decays asymptotically as $e^{-z_v{}^{3/2}}$ for $z > 0$ against the potential of the field, where the particle encounters a triangular barrier starting from the origin. In the direction ($z < 0$), moving with the potential of the field, the wavefunction is that of an accelerating particle and oscillates with increasing frequency as it moves:

$$Ai(z) = \frac{1}{\sqrt{\pi}}\frac{1}{(-z_v)^{1/4}}\sin\left(\frac{2}{3}(-z_v)^{3/2} + \frac{\pi}{4}\right) \tag{10.109}$$

In a semiconductor, both valence band $\vec{k}$-states and conduction band $\vec{k}$-states will turn into Airy functions when a strong field is applied. So the optical admixtures and optical transitions will now be between these new Airy functions labeled c and v rather than the Bloch states considered earlier. In particular it is now possible for a photon to excite any valence band Airy electron state into any conduction band Airy state. Momentum conservation no longer applies because the electrons in a field are not in a well-defined momentum state anymore. Indeed they are constantly accelerated, and this is why the oscillations in shape Eq. (10.109) are getting faster and faster as the electrons move in the direction of decreasing potential energy. The rate at which a charge will be excited from the valence Airy set to the conduction Airy state by the action of a light field is given by Fermi's golden rule:

$$W_{vv'} = \frac{2\pi}{\hbar}\left|\int \Phi_{v,v}{}^*(z) qE_0^z z \Phi_{v',c}(z) dz\right|^2 \delta(E_{v',v} - E_{v,v} - \hbar\omega) \tag{10.110}$$

Here the momentum rule reappears as a reduction in the overlap integral Eq. (10.110) between levels which are not vertically above each other, i.e., differ by the v index of the valence to the v' index of the conduction band Airy states. So we see that non-diagonal transitions $v \neq v'$ are possible, but less likely. Thus a useful quantity for characterizing optical absorption is the local density of states, which, for vertical transitions, is apart from a constant, also the joint density of states discussed before. Remember that the sum of vertical transitions is directly proportional to the joint density of states. The density of states is conveniently expressed using the local density of states which in one dimension is:

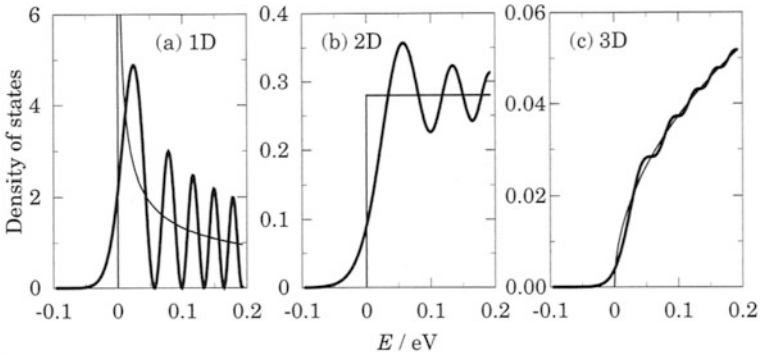


Fig. 10.9 The density of states from left to right for a one-, two-, and three-dimensional free electron system in the presence of an electric field (Davies (1998), p. 211, Fig. 6.2. © Cambridge University Press 1998. Reprinted with the permission of Cambridge University Press)

$$n(E,z) = \sum_n |\Phi_n(z)|^2 \delta(E - E_n) \tag{10.111}$$

where here, Φ_n are the energy eigenstates Eq. (10.108) and the E_n are the eigenvalues. The local density of states say at $z = 0$ gives us a measure of how many eigenstates exist in an energy interval in a given locality. The total density of states *g(E)* is obtained by integrating the local density of states over all space:

$$\int \sum_n |\Phi_n(z)|^2 \delta(E - E_n) dz = \sum_n \delta(E - E_n) = g(E) \tag{10.112}$$

The local density of states, assuming for convenience that the hole electron masses are the same, is a measure of the optical absorption and can be calculated in this case by substituting in the Airy functions Eq. 10.109) and eigenvalues into Eq. (10.112) and doing the sum at $z = 0$. The integrations are straightforward but lengthy. The reader is referred to the details in the books by Chuang (1995) and Davies (1998) for more details. The Franz-Keldysh oscillations in the density of states of free electrons are shown in Fig. 10.9 for a one-dimensional band and also for the two- and three-dimensional systems. Figure 10.9 shows the predicted Franz-Keldysh oscillations in the joint density of states, at the band edge of a semiconductor when an electric field is applied.

When excitons are present in the absorption spectrum, we would expect the electric field to help ionize the excitons and change the absorption spectrum toward the free electron system again. This is indeed observed experimentally at very low temperatures in bulk and at higher temperatures in semiconductor quantum wells as we shall see in Chap. 13 (Fig. 10.10).

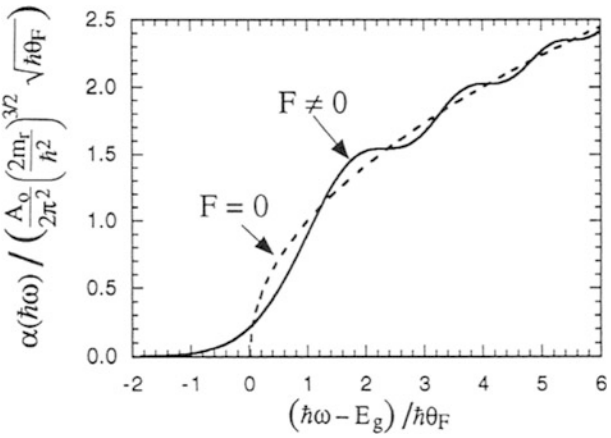


Fig. 10.10 Franz-Keldysh oscillations in the absorption of bulk semiconductors. The dashed line is the spectrum without a field (Chuang (1995), page 549, Fig. 13.2b, Copyright © 1995 by John Wiley & Sons, Inc. Reprinted with permission of Wiley-Liss Inc., a subsidiary of John Wiley & Sons, Inc.)

10.8 Nearly Free Electrons in a Magnetic Field

We now consider the effect of a DC magnetic field on the nearly free electron states of a solid. In order to do this, we write down the Hamiltonian in a field $\vec{B}$ applied in the z-direction. To do this we need to introduce the vector potential $\vec{A}$ and note that in quantum mechanics, the effect of the $\vec{B}$-field is to replace the electron momentum operator $\vec{p}$ with the new operator $(\vec{p} + q\vec{A})$ in the Schrödinger equation. In the so-called Landau gauge, the vector potential is given by $\vec{A} = (0, Bx, 0)$ and $\vec{B}$ becomes:

$$\vec{B} = \vec{\nabla} \times \vec{A} \tag{10.113}$$

and consequently the time-independent Schrödinger becomes:

$$\frac{1}{2m^*}\left[-\hbar^2\frac{\partial^2}{\partial x^2} + \left(-i\hbar\frac{\partial}{\partial y} + qBx\right)^2 - \hbar^2\frac{\partial^2}{\partial z^2}\right]\Psi(x,y,z) = E\Psi(x,y,z) \tag{10.114}$$

From Eq. (10.114) it follows that in the z-direction, the Hamiltonian is that of the free particle, and in the y-direction, the interaction is an x-y product term so we try the solution:

$$\Psi(x,y,z) = u(x)e^{ik_y y}e^{ik_z z} \tag{10.115}$$

with:

$$E = \frac{\hbar^2 k_z^2}{2m^*} + E_{n//} \tag{10.116}$$

Including the spin degree of freedom $s = \pm 1/2$, we also have the Zeeman splitting in a magnetic field:

$$E_{nk_z s} = E_{n//} + \frac{\hbar^2 k_z^2}{2m^*} - sg\mu_B B \tag{10.117}$$

where g is the Lande factor and $\mu_B = \frac{q\hbar}{2m_0}$ is the Bohr magneton.

Substituting the trial function given in Eq. (10.115) into Eq. (10.114), we find that the function $u(x)$ must satisfy:

$$\left[-\frac{\hbar^2}{2m^*}\frac{d^2}{dx^2} + \frac{m^*\omega_c^2}{2}\left(x + \frac{\hbar k_y}{qB}\right)^2\right]u(x) = \epsilon u(x) \tag{10.118}$$

This equation is similar to the one of the harmonic oscillator where:

$$\omega_c = \frac{qB}{m^*} \tag{10.119}$$

is the cyclotron frequency and $\frac{\hbar k_y}{qB}$ is a length which we shall denote with $\{-x_{k_y}\}$. Equation (10.118) is a standard differential equation of mathematical physics which has the Hermite polynomials H_n as solutions. We can therefore now write the complete wavefunction as:

$$\Psi(x,y,z) = Ae^{ik_y y}e^{ik_z z}H_{nk_y}\left(\frac{x - x_{k_y}}{l_B}\right)\exp\left[-\frac{\left(x - x_{k_y}\right)^2}{2l_B^2}\right] \tag{10.120}$$

where n are integers, A is the normalization constant, and $l_B = \sqrt{\frac{\hbar}{qB}}$ is called the magnetic length that is typically ~25 nm for $B = 1$ T. The first few normalized Hermite polynomials are tabulated and given by:

$$\begin{aligned} H_0(s) &= 1 \\ H_1(s) &= 2s \\ H_2(s) &= 4s^2 - 2 \end{aligned} \tag{10.121}$$

The corresponding x-y energy levels are independent of the index k_y and given by (n is an integer including 0):

$$E_{//} = \varepsilon_n = (n + 1/2)\hbar\omega_c \tag{10.122}$$

These levels are called the Landau levels. Each Landau level is highly degenerated because there are many k_y levels in each Landau level. In fact there are exactly $\frac{L_x L_y qB}{h}$ states in each Landau level, apart from spin which is another factor 2. Thus the degeneracy grows with B because the separation of the levels also grows with B. When we include the spin, the Landau spin up and spin down bands are shifted relative to each other by the Zeeman energy $g\mu_B B$. The collapse of the x-y spectrum into discrete Landau levels is a novel phenomenon with strong consequences for the transport and optical properties of systems with free carriers. The condition for observing subtle effects in transport and optical spectra which are caused by the magnetic field is that the energy levels should have long relaxation times, so that the broadening of the levels should satisfy the condition that $\frac{\hbar}{\tau} < \hbar\omega_c$. This condition is difficult to satisfy in practice because in a metal $\tau \sim 10^{-13} - 10^{-14}$, which gives a much larger uncertainty $\Delta E \sim \hbar/\tau$ than the typical Landau level separation which is $\hbar\omega_c \sim 10^{-4} eV$ at $B = 1$ T. To observe the effect of Landau levels experimentally, one has to work with very high-quality and low effective mass semiconducting materials and preferably quantum wells that are systems composed of a thin lower bandgap semiconductor layer sandwiched between two higher bandgap materials (see Chap. 15 for details).

Normally, one also has to work at very low temperatures. Good materials for large Landau level separations are, for example, GaAs and InAs and InSb which would enhance the $B = 1$ T splitting by a factor 40 (InAs: $m_e^*/m_0 = 0.023$) to 4×10^{-3} eV or 70 (InSb: $m_e^*/m_0 = 0.0145$) to 7×10^{-3} eV which is ~70 K. We will come back to this topic when we discuss the low-dimensional semiconducting systems in Chap. 15. As before, the easiest way to study the effect of Landau levels on optical absorption theoretically is to evaluate the local density of states by substituting the wavefunctions and energies into Eq. (10.111) and carry out the sum.

In a two-dimensional system, for example, which one can engineer with a quantum well structure, the free electron density of states can be computed in the same way as we did for the three-dimensional case (Chap. 5 replace $4\pi k^2 dk \rightarrow 2\pi k dk$ in Eq. (5.37)). It is constant for the two-dimensional case and given by $g_2(E) = \frac{Sm^*}{\pi\hbar^2}$ where S is the area of the system. When subject to a B field, we see from the above solution that we now only have the Landau spectrum, and the Landau level density of states now consists of sharp delta function peaks for each Landau level. The sharp delta function peaks are of course unrealistic, and one has to evaluate the sum by including a finite level broadening before plotting the function.

Figure 10.11 illustrates how the two-dimensional constant density of states collapses into Landau levels which are not ultrasharp delta functions but broadened by disorder or phonon scattering processes. Thus in a two-dimensional system, the electrons would fill the Landau levels up to the Fermi energy. The Fermi energy can then be in the Landau band or in a gap, depending on the electron concentration and the magnitude of the magnetic field. Such quasi-two-dimensional systems can be made using multilayers and quantum wells as we shall see in detail in Chap. 15.

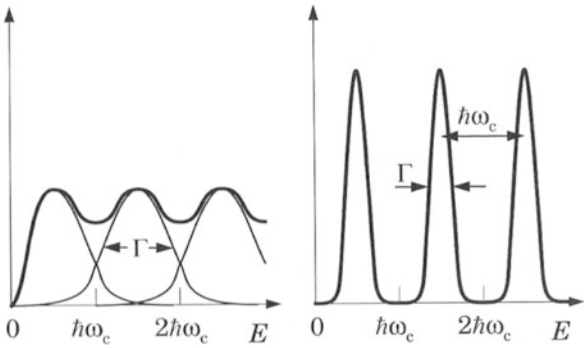


Fig. 10.11 The density of states of a two-dimensional electron gas in a magnetic field for two different values of broadening. As the broadening is reduced, the Landau levels become delta function like peaks. With increased broadening, the trend is to a constant density of states as in the $B = 0$ limit (Davies (1998), p. 225, Fig. 6.7b and 6.7c. © Cambridge University Press 1998. Reprinted with the permission of Cambridge University Press)

By changing the magnetic field, it is therefore possible to move the Fermi energy inside the Landau bands, and from inside the band to the gap between adjacent bands, when the bands are full. In Eq. (10.40) we made the observation that when the density of states at the Fermi level is zero, there is no conduction. By changing the magnetic field, it is therefore possible to make the two-dimensional system undergo a transition from a conducting to a nonconducting state. This happens because by changing the level density in each Landau subband, one can move the Fermi level from inside a Landau band to a gap. Thus the resistance of a two-dimensional gas is expected to oscillate with magnetic field, a phenomenon known as Shubnikov de-Haas effect, and this is indeed observed in high-quality semiconducting quantum wells. This is discussed in more detail in Chap. 15.

In a three-dimensional system, the k_z degree of freedom broadens the Landau bands, and we have (spinless case):

$$g(E) = \frac{qBL_xL_y}{h} \sum_{n,k_z} \delta\left(E - \varepsilon_n - \frac{\hbar^2 k_z^2}{2m^*}\right) \tag{10.123}$$

$$g(E) = \frac{qVB\sqrt{2m^*}}{(2\pi\hbar)^2} \sum_{n=0}^{n_{\max}} [E - (n + 1/2)\hbar\omega_c]^{-1/2} \tag{10.124}$$

where $n_{\max}$is the highest allowed subband index below the given energy E.

The conductivity is included in the total permittivity, so the magnetic field can in principle strongly change the refractive index of the system. The key factor in magneto-optics is however the broadening, which is, as we have seen, in most systems, larger than the Landau level separation. In practice one cannot go to fields

much higher than about 17 T, and this therefore severely limits possible technical applications of orbital magnetism to optoelectronics.

The permittivity of free electrons in a magnetic field can be computed using the wavefunctions we obtained Eq. (10.120) and substituting them into Eq. (10.11); however, it is often adequate to compute the optical spectra of materials within a semiclassical treatment. This can be done by adding the Lorentz force $F_L = -q\dfrac{d\vec{r}}{dt} \times \vec{B}$ to the right-hand side of Eq. (10.69), the Newton equation of motion, and evaluating the magneto-Drude polarization response just as we did before. With the B field in z-direction, the Lorentz force makes the problem necessarily two-dimensional in the x-y plane, because it introduces a transverse Hall velocity, so that now we have two equations for the two velocities v_x and v_y in response to the x-electric field. Assuming that the light vector is polarized in the x-direction as in Eq. (10.18), we can solve for the permittivity as we did before, but now including the Lorentz force and neglecting the phonon contribution, we find from Eq. (10.26):

$$\begin{aligned} m^* \frac{d^2x}{dt^2} + m^* \frac{dx1}{dt\,\tau} &= -qE(t) - q\frac{dy}{dt}B \\ m^* \frac{d^2y}{dt^2} + m^* \frac{dy1}{dt\,\tau} &= q\frac{dx}{dt}B \end{aligned} \tag{10.125}$$

These equations are solved by making the same assumption as before for the displacements $x(t) = x_0 e^{-i\omega t}$ and $y(t) = y_0 e^{-i\omega t}$. We find the new B field-dependent free carrier relative permittivity contribution and add it to the bound relative permittivity to obtain:

$$\varepsilon(\omega) = \varepsilon_b(\omega) + \frac{i}{\omega\varepsilon_0}\sigma(B,\omega) \tag{10.126}$$

where the complex conductivity now is dependent on the B field via the cyclotron frequency:

$$\sigma(B,\omega) = \frac{n_c q^2 \tau}{m^*}\left(\frac{1}{\tau}\right)\left\{\frac{1/\tau - i\omega}{(i\omega - 1/\tau)^2 + \omega_c^2}\right\} \tag{10.127}$$

Equation (10.127) reduces to the usual result Eq. (10.30) when the magnetic field B goes to 0.

Since absorption is related to the imaginary part of the permittivity, and the bound term can be treated as real for frequencies below 10^{13} Hz, the absorption coefficient is proportional to the real part of the conductivity. Indeed we have from Eqs. (10.130 and 10.15):

$$\alpha(\omega) = \frac{\omega}{c}\left\{\frac{\left((1/\tau)^2 + \omega_c^2 + \omega^2\right)}{\left[(1/\tau)^2 + \omega_c^2 - \omega^2\right]^2 + 4\omega^2/\tau^2}\right\}\frac{n_c q^2}{\varepsilon_0 m^*}\left(\frac{1}{\omega\tau}\right) \tag{10.128}$$

The absorption exhibits resonance absorption at light frequencies which match the cyclotron frequency ω_c shifted by the relaxation broadening. This resonance is called the cyclotron resonance and is most important for measuring the cyclotron frequency or what in other words is the effective mass of the electrons. The resonance can be understood immediately in the quantum mechanical picture as the absorption of a photon when an electron goes from one Landau level to the next. The semiclassical result suggests that most of the oscillator strength is indeed associated with a transition from one to the next adjacent Landau level, as is the case in the harmonic oscillator problem.

The full quantum mechanical treatment of magneto-optics is very rich in information. The formalism gives rise to complex expressions which are sometimes difficult to handle analytically. The full treatment is normally not necessary unless one is truly in the limit of long coherence lengths, or small broadening, i.e., broadening smaller than the Landau level spacing. This is achievable with very high-quality semiconductors at low temperatures, but almost never in a metal. Figure 10.12 shows the change in the optical absorption edge of InSb caused by a magnetic field. The reader should also refer to the discussion presented in Chap. 15.

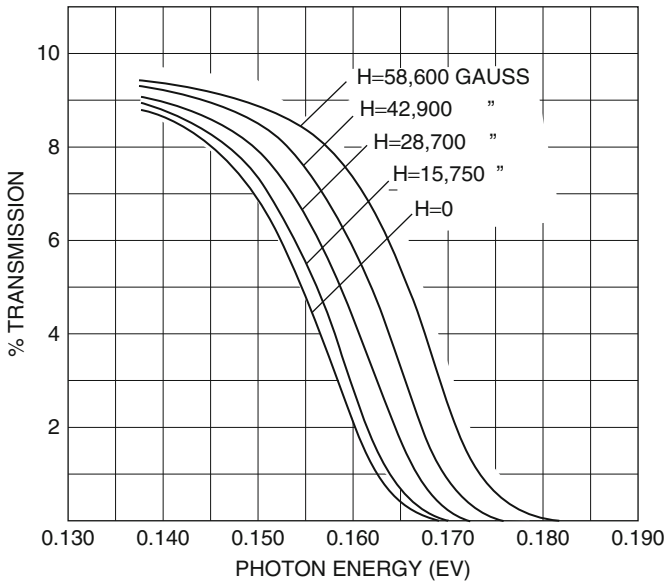


Fig. 10.12 The band-edge absorption of InSb with magnetic field at room temperature (Reprinted figure with permission from Burstein et al. (1956), p. 827, Fig. 1. Copyright 1956 by the American Physical Society)

10.9 Nonlinear Optical Susceptibility

We have seen how a medium affects light and how this can be described by the concept of permittivity and complex refractive index. Throughout however, we assumed that the light wave constituted only a weak perturbation on the electronic and lattice coordinates. It was therefore sufficient to allow the light vector to couple with these modes and consider the response of these modes to first order in the light electric field. The dipole moment that the light induced was evaluated in linear response only. Even though we did allow other external electric and magnetic fields of arbitrary magnitude to act on the system, this was not the electric field of the photon. One may therefore ask: what happens when the photonic field is so strong that higher-order processes in the optical permittivity or susceptibility become important? The first thing we note is that in this case, we need to compute the polarization $\vec{P}$ to higher orders in the electric light field $\vec{E}$, so we write in the usual tensor notation:

$$\vec{P} = \chi^{(1)}\,\vec{E} + \chi^{(2)}\,\vec{E}\cdot\vec{E} + \chi^{(3)}\,\vec{E}\cdot\vec{E}\cdot\vec{E} + \ldots \tag{10.129}$$

or equivalently:

$$P_i = \sum_l \chi^{(1)}_{il} E_l + \sum_{l,k} \chi^{(2)}_{ilk} E_l E_k + \sum_{lks} \chi^{(3)}_{ilks} E_l E_k E_s + \ldots \tag{10.130}$$

where $\chi^{(n)}$ are the susceptibility tensors.

When the field is time dependent, the susceptibilities can be evaluated by the same method as used in Eq. (10.43) for the first-order term, i.e., by using time-dependent perturbation theory and going to higher orders. When the electric field frequency is not monochromatic, i.e., if $E(t) = \sum_{\omega_\mu} E_\mu e^{i\omega_\mu t}$, the susceptibilities will depend on two frequencies for the second-order term, on three frequencies for the third-order term, etc., and the sums in Eq. (10.129) will run over frequencies as well.

The physical significance of the higher-order terms will now be explained. The first-order term contains the one-photon absorption or emission process which is what we have discussed until now, having specialized the analysis to a polarized electric field and the term $\chi_{xx}(\omega) = \alpha_p(\omega)$ only. Similarly, the second-order term describes processes which allow two photons to be absorbed or emitted simultaneously. It includes also the process in which a photon is converted into a lower- or higher-energy one (with phonon absorption or emission). The second-order term only exists in crystals which have no center of inversion symmetry. When they do, then this term vanishes by symmetry. The third-order term is always there, but the second-order term can sometimes be induced by applying a strong additional static external field which breaks the symmetry of the crystal. The third-order term involves three-photon processes, for example, two absorbed and one emitted or vice versa. It is clearly highly desirable to be able to do that kind of

photon-to-photon energy conversion with high efficiency and reproducibly many times over. Unfortunately, the higher-order susceptibilities get progressively weaker with order, and such conversions are normally inefficient and require high laser power. The high laser power then damages the material with time, and this constitutes a serious problem. The field of nonlinear optics is therefore very well developed. Many materials including organic and inorganic ones have been studied, and the reader is referred to the specialized literature on the subject (Peyghambarian et al. 1993).

Let us return to the first-order term in the above expansion and now allow an external say static field to modify the permittivity. This is a most important scenario and gives rise to the so-called electro-optic and magneto-optic effect. It allows us to change the complex permittivity of a medium by applying an external field. The basic theory for evaluating the electro- and magneto-optical effects was developed above. The "ease" with which a medium changes its permittivity under the action of such a field is measured by the so-called electro-optic coefficients. These can be obtained as the coefficients of the expansion of the permittivity with the external applied fields. The refractive index of materials such as $LiNbO_3$ (one of the best), KH_2PO_4, and even GaAs responds relatively strongly to an applied electric field. In the material which we are familiar with, namely, GaAs, the applied electric field will, for example, change the band structure and bandgap by replacing the Bloch states with Airy functions and in this way give rise to a new refractive index. This refractive index can be calculated by first evaluating the field-dependent polarization using the above formalism. For more details and a more quantitative analysis, the reader is referred to the book by Chuang (1995).

10.10 Summary

In this chapter we have presented a detailed and reasonably complete treatment of the optical permittivity of a solid. We have shown how one can relate absorption, refraction, reflection, and transmission of light to the real and imaginary parts of the complex refractive index. Then we showed how the refractive index has to be computed in different types of solids. We started with the free electron contribution, then added the bound electrons, and finally included the photon-phonon coupling. Only polar optical phonon modes were included, which of course covers only a very small part of the field. It was shown how photon-phonon coupling can lead to the formation of a new type of particle called the polariton. The polariton is "part photon part phonon" and is a very beautiful effect. We also mentioned, but did not develop, the science of the surface plasmon. We showed how absorption could be related to quantum transitions. For this we had to derive the important rule called Fermi's golden rule which gives us the rate of transfer from one eigenstate to another under the action of a time-dependent perturbation. We specialized the permittivity calculation to the case of semiconductors and introduced a very elegant way of computing the Bloch matrix elements known as the Kane parameter method derived from the Kane effective mass method.

We introduced the reader to the quantum mechanics of nearly free electrons subjected to the effect of strong electric and magnetic fields. The corresponding Franz Keldysh and Landau wavefunctions and energy levels were derived, and we showed how electric and magnetic fields changed the density of states of electrons. The new quantum energy spectra affect both transport properties and optics, but these are highly specialized themes which need detailed focused treatment. We introduced the reader to the fundamental new science, the new concepts, and the methodology needed to compute the optical permittivities with some simple examples. Magnetic and electric fields can be very effective tools for the modulation of optical properties, with a strong impact on technology. This is especially true in low-dimensional systems, so we defer the discussion on some of the applications to the chapter on low-dimensional solids. One problem is that magnetic and electric fields, however weak, can never be treated mathematically in perturbation theory using the unperturbed Schrödinger equation, when we have an infinite unbounded system. The magnetic perturbation involves a term $\sim -x^2$ which binds electrons in one direction and the electric field a term $\sim -qE_0^z z$ which is unbounded as $z \to \infty$. The quantum treatment can be technically tedious because we are forced to use the exact wavefunctions derived above; however, it weaks the perturbation. These exact wavefunctions are, as one can verify, not at all simply related to the free electron-like waves. In this context it is therefore noteworthy that the semiclassical methods, when applicable, can be very useful. This was shown here in the magneto-optical example. In finite quantum confined systems on the other hand, the wavefunctions are bounded and normalized in a finite volume. Here one can treat electric and magnetic fields using second-order perturbation theory and get good results. This then also allows one to evaluate the electro-optic coefficients using perturbation theory. We shall look at this in more detail in Chap. 15.

Problems

1. Calculate the real and imaginary part of the frequency-dependent admittance of a wire as a function of frequency, if the area is 1 cm^2, the length 0.1 cm, the charge density 10^{21} cm^{-3}, and the relaxation time $\tau = 10^{-13}$s and effective mass $0.1m_0$. Write down the results as a function of frequency. What are the conductance and the capacitance?
2. Calculate the oscillator strength F_{12} linking the ground state $n = 1$ and first excited state $n = 2$ of box eigenstates with box size $L = 1$ nm and effective mass $m^* = 0.023\ m_0$.
3. Calculate the reflectivity of a metal as a function of frequency using the Drude permittivity formula with free carrier concentration $n_c = 10^{22}$ cm^{-3}, relaxation time $\tau = 10^{-12}$ s, and $m^* = 0.045\ m_0$. Plot the result and compare with Fig. 10.2.
4. Explain the difference between direct and indirect bandgap materials. Sketch the two situations. If phonons were not allowed to provide the necessary momentum

in an indirect bandgap excitation, what other mechanisms can you think of which could make the absorption process happen in another way?

5. Calculate the density of states per unit volume of a three-dimensional nearly free electron gas with effective mass m^* in a magnetic field B_z perpendicular to the x-y plane including spin. Remember that the number of allowed k_y states per Landau level is given byL_xL_yqB/h for an area of size L_xL_z and that there is another (free electron) z-degree of freedom in the z-direction.
6. What is meant by the permittivity of a solid? How is it calculated? How is it related to the refractive index? What does the real and imaginary part of the refractive index signify? How would you design a material which is a perfect reflector?
7. Using the definition of the complex refractive index given by Eq. (10.9), derive the pair of equations given by Eq. (10.14) which show that this leads to a quadratic equation from which the real and imaginary part of the complex refractive index $\bar{n}$ and κ can be computed.
8. What is a phonon-polariton? Write down the explicit algebraic solutions which give the two branches of the dispersion relation $\omega^2(k)$ for the phonon-polariton equation using Eq. (10.102). Explain how and why the group velocity of this new particle changes with wavenumber.
9. What is an exciton? In GaAs the effective mass of an electron is $m_e = 0.067\, m_0$, and the effective mass of the hole is $m_h = 0.082 m_0$. The relative static permittivity ε_r is 13.1. Using Eq. (10.85) and Eq. (10.86), calculate the exciton radius and binding energy. At what temperatures would you expect the excitons to be detectable by experiment?
10. With the help of Eq. (10.125), derive the magnetic field-dependent complex conductivity of an electron gas as given by Eq. (10.127): $\sigma(B,\omega) = \dfrac{nq^2\tau}{m^*}\left(\dfrac{1}{\tau}\right)\left\{\dfrac{1/\tau - i\omega}{(i\omega - 1/\tau)^2 + \omega_c^2}\right\}$. Discuss the behavior of the real part as a function of the magnetic field. What happens when the magnetic field becomes very large? Give a physical interpretation. How does a magnetic field affect the reflectivity of a free electron gas?

References

Burstein E, Picus GS, Gebbie HA, Blatt F (1956) Magnetic optical bandgap effects in InSb. Phys Rev 103:826–828

Chuang SL (1995) Physics of optoelectronic devices. Wiley, New York

Davies JH (1998) The physics of low dimensional semiconductors: an introduction. Cambridge University Press, Cambridge

Newman R, Tyler WW (1959) Photoconductivity in germanium. In: Seitz F, Turnbull D (eds) *Solid state physics*, vol 8. Academic Press, New York, pp 49–103

Peyghambarian N, Koch SW, Mysyrowicz A (1993) Introduction to semiconductor optics. Prentice-Hall, Englewood cliff

Rosencher E, Vinter B (2002) Optoelectronics. Cambridge University Press, Cambridge

Seeger K (1997) Semiconductor physics: an introduction. Springer, Berlin

Sturge MD (1962) Optical absorption of Gallium Arsenide between 0.6 and 2.75 eV. Phys Rev 127:768–773
Turner WJ, Reese WE (1962) Infrared lattice absorption of AlSb. Phys Rev 127:126–131

Further Reading

Bockrath M, Cobden DH, Lu J, Rinzler AG, Smalley RE, Balents L, McEuen PL (1999) Luttinger liquid behaviour in carbon nanotubes. Nature 397:598–601
Bastard G (1988) Wave mechanics applied to semiconductor heterostructures. Halsted Press, New York
Cohen-Tannoudji C, Diu B, Laloë F (1977) Quantum mechanics. Wiley, New York
Davydov AS (1965) Quantum mechanics. Pergamon, New York
Kittel C (1976) Introduction to solid state physics. Wiley, New York
Liboff RL (1998) Introductory quantum mechanics. Addison-Wesley, Reading
Madelung O (1978) Introduction to solid state theory. Springer, New York
Powell JL, Crasemann B (1961) Quantum mechanics. Addison-Wesley, Reading
Ziman JM (1964) Principles of the theory of solids. Cambridge University Press, London
Ziman JM (1969) Elements of advanced quantum theory. Cambridge University Press, London

Solar Energy Harvesting

11

11.1 Photovoltaic Cells (PVC) Introduction

The sun is a serious and vital source of energy, without which there would be no life on the planet. Plants get most of their energy from the sun by a process called photosynthesis (Jordan et al. 2001). Though fascinating and beautiful, the mechanism of photosynthesis is beyond the scope of this book, and the interested reader is advised to follow up the vast literature on this subject which encompasses physics, chemistry, and biology. But a part of the sun's energy can also be harvested artificially using photoconductive devices (Pohlman, Heeger). This subject and technology has become of supreme importance since the realization that fossils fuels are slowly but surely destroying our planet. The sun emits light over a broad spectrum of frequencies as shown in Fig. 11.1.

The process of photon harvesting is illustrated in Fig. 11.2, and one can see that semiconducting p-n junctions (Chap. 6, this book) are the ideal way to collect the photonic power. But there are restrictions here too. One can see from the diagrams that the photon energy must exceed the bandgap of the material to be absorbed efficiently. So depending on the semiconductor in question, "all" the photons above the bandgap can be harvested, but this also means that the solar photons below the bandgap are not harvested. The latter constitutes in general a non-negligible amount. This implies that semiconducting solar cells are not as efficient as they could be if they collected the entire spectrum. Si or GaAs, which are some of the best *PVCs*, leave out the photons below 1 eV (> 1200 nm), and this is an important loss limiting efficiency to ~20%. Indeed, combination cells which are designed to collect a wider range of wavelengths can nowadays reach efficiencies of 45% (see below); the problem is that they are still too expensive for large-scale commercial application.

Long-wavelength collection can be done with type II semiconducting devices (Delaunay et al. 2008), which are also used for long-wavelength photodetection.

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_11

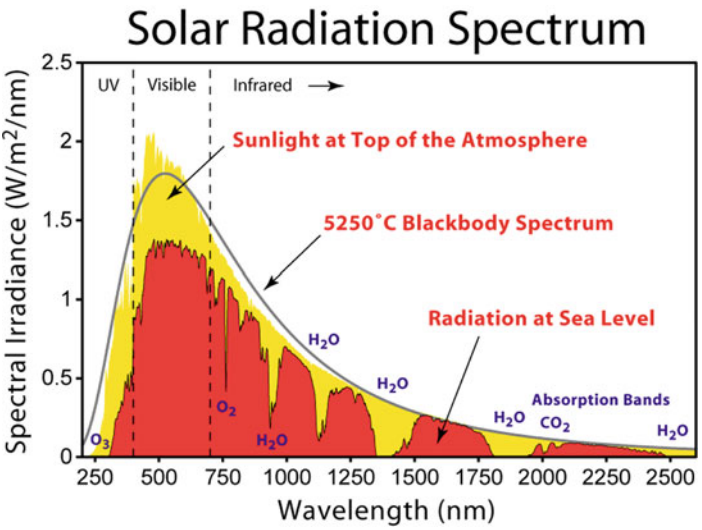


Fig. 11.1 The most important range is in the UV to visible to near-infrared range (300–1000 nm). The total solar power reaching the surface is roughly 1000 W/m^2 on average, a non-negligible amount. The mentioned photon range is ideal for the application of semiconducting p-n junction technology. Remember Fig 9.2 from Chapter 9

Figure 11.3 is a beautiful illustration of a device which can be used for long wavelength (>3 μm; the same geometry is used for making top of the scale photodetectors operating currently at 200 K).

11.2 Examples of Photodiodes

*For commercial use of *PVC* devices, efficiency is not the only criterion. Many applications require mechanical flexibility and thus polymer cells or biocompatibility (plastic electronics implants into the body) (Figs. 11.4, 11.5, and 11.6).

11.3 The Current Voltage Characteristic of a Solar Cell (Figs. 11.7 and 11.8)

$$
\begin{aligned}
L &= \text{width} \\
v_\mathrm{d} &= \text{drift-velocity} \\
\tau &= \text{recombination-time}
\end{aligned}
$$

$$
\eta = \frac{v_\mathrm{d}\tau}{L}\left[1 - \exp[-L/v_\mathrm{d}\tau]\right] \tag{11.1}
$$

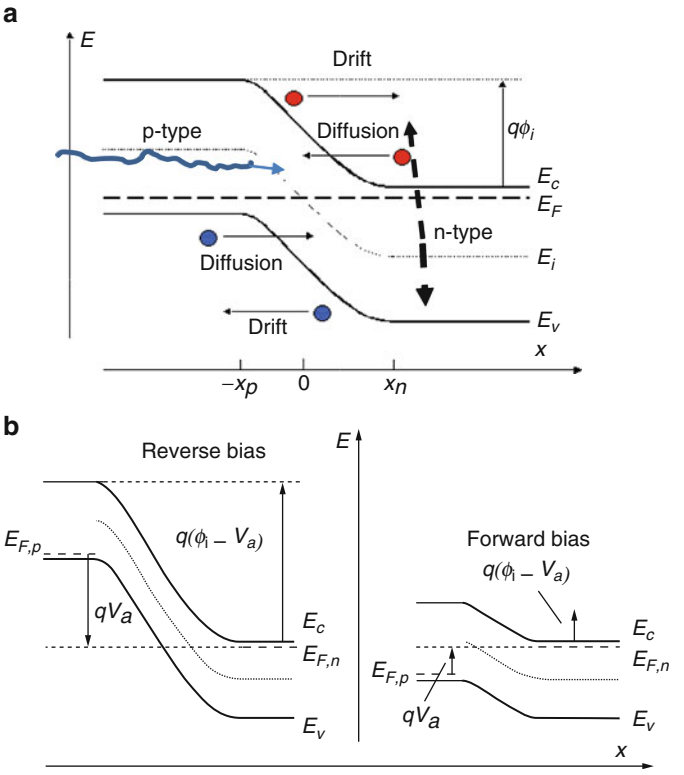


Fig. 11.2 (**a**) A photon (blue wiggly line) comes into the high-field region and excites an electron (red) hole (blue dot) pair across the gap of a p-n junction semiconductor. The electron and hole are subject to an internal space charge field which makes them drift into the electrode regions where they are absorbed, thus creating a current without an applied bias. This current can heat a resistor in series and thus constitutes harvested solar energy. (**b**) Showing (left) a semiconductor junction under reverse and (right) forward bias. Note that reverse biasing enhances the internal field and facilitates charge collection

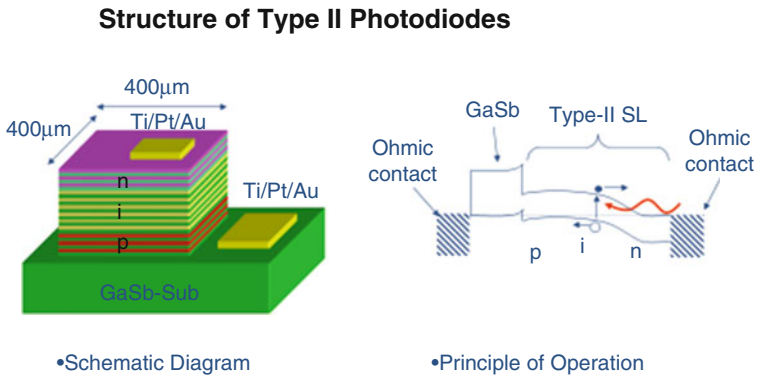


Fig. 11.3 Type II photon detection or light-harvesting structure (Delaunay et al. 2008)

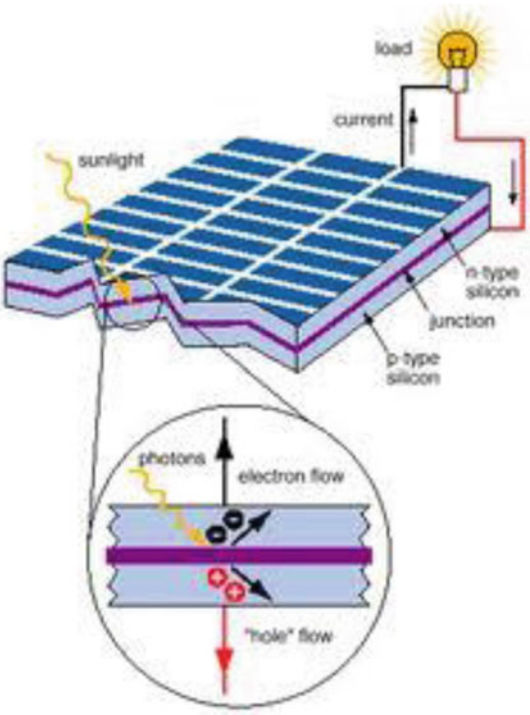


Fig. 11.4 Schematic representation of a photo-harvesting device

Let us now consider the limit in which charges generated in the cell can only drift and are collected at the perfect absorbing electrode, or they recombine in the material with lifetime τ.

$$\eta = \text{quantum} - \text{efficiency} =$$

$$\text{charge} - \text{collected} - \text{per} - \text{photogenerated} - \text{charge}$$

$$\eta = \frac{v_d \tau}{L}[1 - \exp[-L/v_d\tau]] \qquad (1)$$

$$L = \text{width}$$

$$v_d = \text{drift} - \text{velocity}$$

$$\tau = \text{recombination} - \text{time}$$

Equation 11.1 is the expression for the *QE (quantum efficiency)* η in the drift limit (no back diffusion) (called the *Hecht formula*).

11.3.1 Solar Cell IV Characteristic Curve

It explains how much power can be extracted for a given photogenerated current IV.

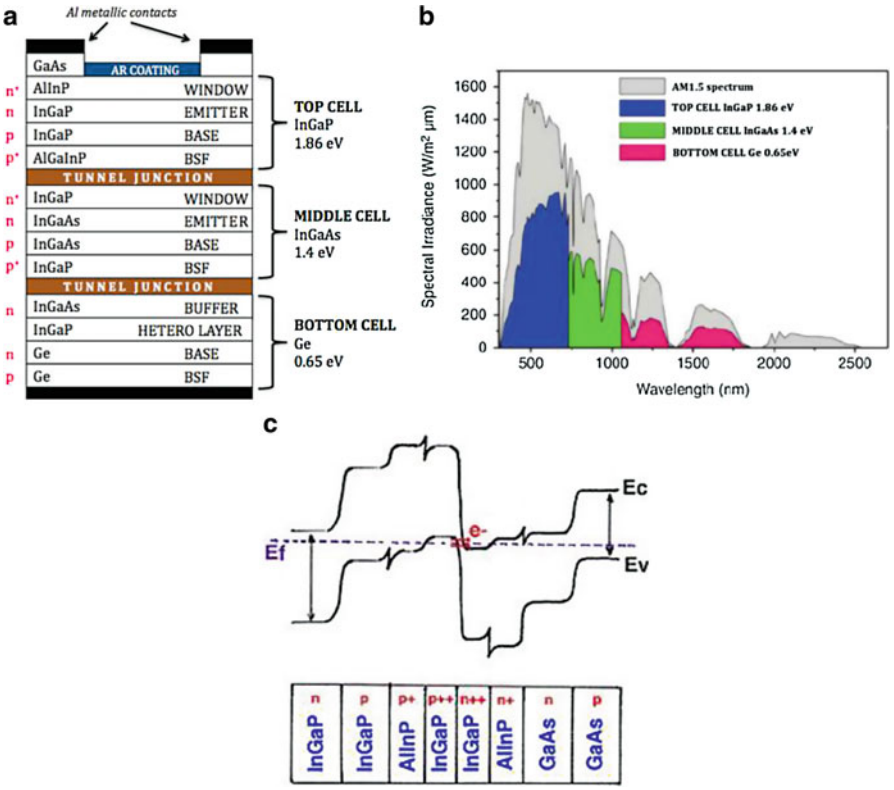


Fig. 11.5 (**a**) (top figures = example of *multijunction cell* structure (left) and harvesting ranges (right), lower figure corresponding band structure) Combination cells which harvest a wider region of the sun's spectrum. Note how the interfacial barrier is designed to be thin enough for carriers to tunnel through, lower figure. (**b**) Band structure of a multijunction cell showing the way the two modules work together. Electrons generated in one module recombine with holes through thin barrier in the other to produce a current (Yamaguchi et al. 2005)

11.4 General Expression for the Quantum Efficiency

Let us consider the semiconductor channel in a p-n junction and model it as a one-dimensional system since the planar motion is uniform. Light impinges from the left a shown in Fig. 11.10 and creates e-h pairs which drift diffuse into the electrodes. This time the electrodes are not considered a being totally absorbing, but they have finite surface recombination velocities s_1 and s_2. The width of the depletion layer is taken as w and $t + w$ is the total length.

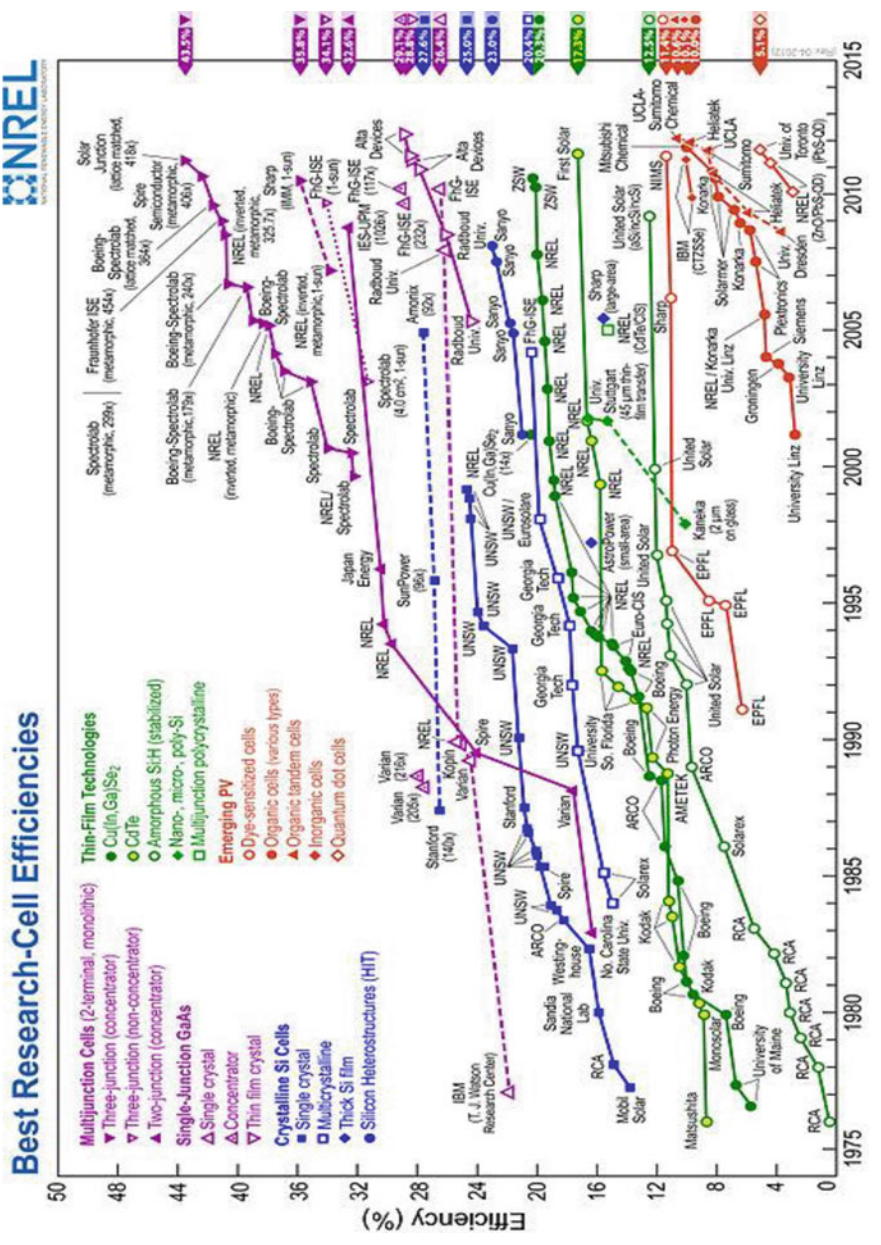


Fig. 11.6 Map of best current solar cell efficiencies in all technologies; this map has to be updated every 6 months

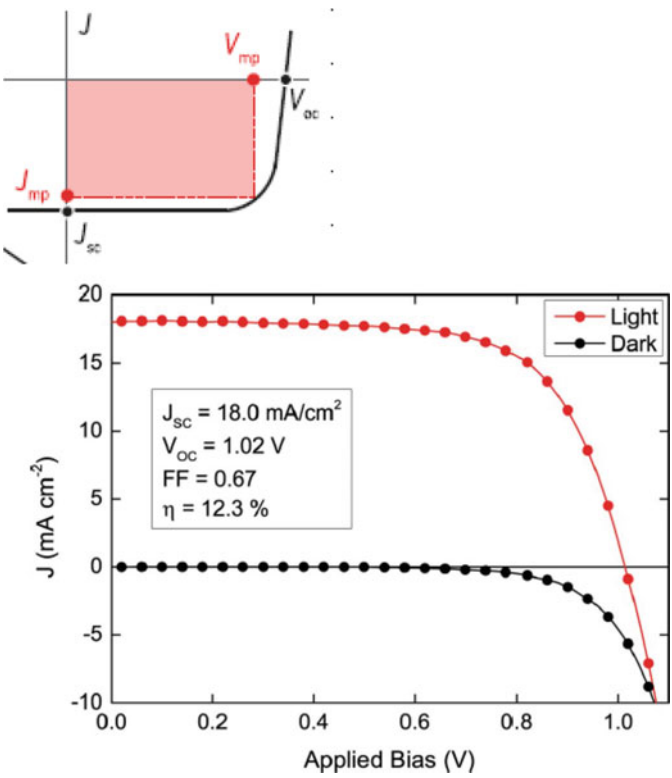


Fig. 11.7 Current voltage characteristic under simulated light of 1.5 suns or 100mw/cm^2 solar irradiation of the best performing perovskite solar cells ($\eta = 12.3\%$) (From Bail JM, Leed M, Hey A, Henry J Snaith Energy Envir. Sc. Vol.6, p 1739 (2013) "Low-temperature processed meso-superstructured to thin-film perovskite solar cells")

11.5 Some Definitions, Power Collected

Consider Λ = power collection efficiency (not to be confused with quantum efficiency for charge collection η); V_{oc} = *open circuit voltage*, or forward bias at which the bias produced current cancels photocurrent. FF = fill factor of IV curve, deviation of IV curve from perfect rectangular shape; I_{sc} maximum photocurrent at zero bias; see Fig. 11.9.

$$\Lambda = V_{oc} I_{sc} FF / P_{in} \tag{11.2}$$

$$FF = \text{Fill} - \text{factor} \tag{11.3}$$

$$I_{sc} = \text{sat} - \text{current} \tag{11.4}$$

$$P_{in} = \text{total} - \text{power} - \text{incident} \tag{11.5}$$

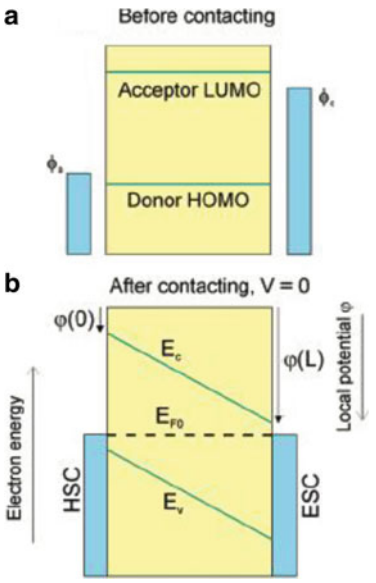


Fig. 11.8 Top diagram, simple semiconductor bands connected to two electrodes where the Fermi energies are roughly matched to the conduction and valence band, respectively, and diagram below the band alignment that follows on contact

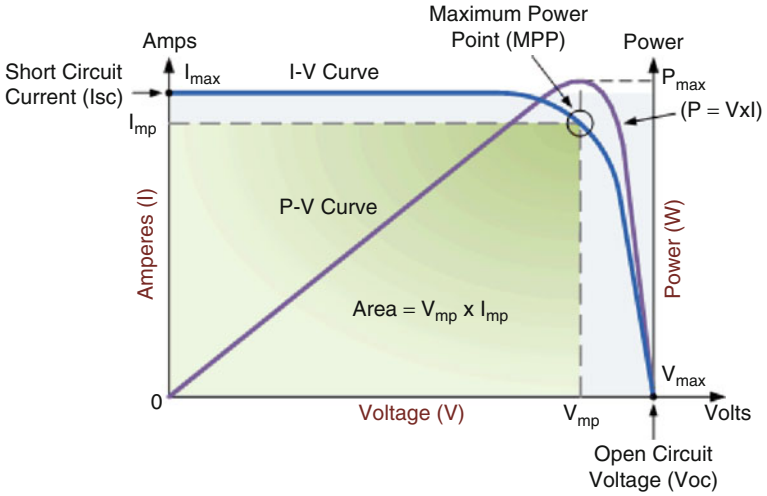


Fig. 11.9 Illustration of the ideal situation with corresponding definitions (Credgington & Durrant 2012)

In the limit of no back diffusion, all carriers generated by light drift into the electrode or recombine in the bulk unless they recombine in the bulk with lifetime τ. From Fig. 11.10 which explains what is meant by power collection efficiency, we note that if a semiconductor has a low bandgap and thus high dark current at room temperature, then the V_{oc} is small and the IV area is reduced and thus of lower efficiency. Let us now cone the complete formula for the quantum efficiency η in a p-n junction such as in Fig. 11.1. The *saturation current* is just the product of the number of photons absorbed and efficiency η.

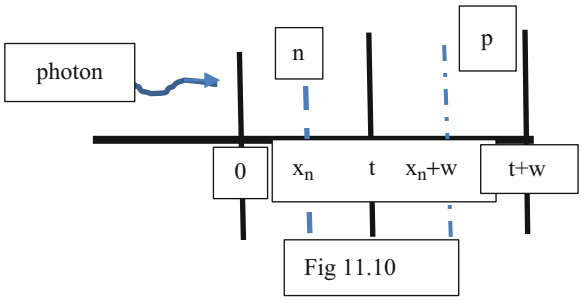


Fig. 11.10 Rosencher & de Vinter (2005)

11.6 Complete Mathematical Expression for the Quantum Efficiency

The quantum efficiency (charge collected per charge created) η is divided in three contributions: the n-region, the high-field region DR, and the p-region.

The theoretical evaluation, which involves solving the diffusion equation with boundary conditions, is given below noting that (Movaghar & Schirmacher 1981):

r = reflection coefficient, α = absorption coefficient, and $\gamma_1 = s_1 L_h/D_h$, $\gamma_2 = s_2 L_e/D_e$.

L_e and L_h are the electron and hole diffusion length, respectively. D_e and D_h are the electron and hole diffusion coefficient; *s_1 and s_2 are the surface recombination* velocities at the illuminated and back photodiode surface, respectively.

$$\eta = \eta_e + \eta_{DR} + \eta_p$$

Respectively, n-region, high-field region DR, and p-region:

$$\eta_p = \frac{(1-r)\alpha L_e}{\alpha^2 L_e^2 - 1} e^{-\alpha(x_n+w)} S \tag{11.7}$$

$$S = \frac{(\gamma_2 - \alpha L_e)e^{-\alpha(t+d-x_n-w)} - sh[(t+d-x_n-w)/L_e] - \gamma_2 ch[(t+d-x_n-w)/L_e]}{ch[(t+d-x_n-w)/L_e] + \gamma_2 sh[(t+d-x_n-w)/L_e]} + \alpha L_e \tag{11.8}$$

$$\eta_n = \frac{(1-r)\alpha L_h}{\alpha^2 L_h^2 - 1}\left(\left[\frac{\alpha L_h + \gamma_1 - e^{-\alpha x_n}(\gamma_1 ch(x_n/L_h) + sh(x_n/L_h))}{\gamma_1 sh(x_n/L_h) + ch(x_n/L_h)}\right] - \alpha L_h e^{-\alpha x_n}\right) \tag{11.9}$$

$$\eta_{DR} = (1-r)\left[e^{-\alpha x_n} - e^{-\alpha(x_n+w)}\right] \tag{11.10}$$

$$\gamma_1 = s_1 L_h/D_h, \quad \gamma_2 = s_2 L_e/D_e \tag{11.11}$$

The regions x_n, t, and w are defined in the figure; see *A Rogalski ad J Rutkowski Infrared Phys Vol 22 p199 (1982).*

This is the complete mathematical expression for the number of carriers collected per photon when we allow for both diffusion and drift and the fact that the carriers are not necessarily collected with unit absorbing efficiency when they reach the electrode boundaries, but that there can be back reflection expressed by the finiteness of the so-called surface recombination velocities s_1 and s_2. Note that there are three regions of charge generation n,p and depletion region DR. The simple Eq. (11.1) called also the Hecht formula (Hecht 1932) is recovered in the limit that the γ's are infinite, and we only have one carrier type $r = 0$ and only n-region.

11.7 Summary: Discussion

We have given in this chapter a brief description of the semiconducting solar cell used for light harvesting. The focus is on inorganic materials. The subject in general is a vast one with colossal importance to society in view of the gigantic damage being caused by fossil fuels and global warming. Work is going on all over the world in trying to still raise the efficiencies of solar cells. Electric cars and trains will be the dominant form of transport for sure, and the target of getting rid of fossil fuels is now very near. The combination devices illustrated in Fig. 11.5 exhibit great values, and, were it not for the high production cost and maintenance charges, one could consider the problem as almost solved. The difficulty is to harvest a larger part of the solar radiation than, for example, silicon or GaAs (>1 eV), without too many expensive processing and manufacturing steps. Scientists are working to solve this problem by tackling the problem from many directions: new materials and new geometries. We also reported on the important breakthrough made recently in making polymer solar cells such as the system *P3HT/PCBM* (Polman et al. 2016; Street & Schoendorf 2010; Sariciftci et al. 1992) which have reached power efficiencies η_{sp} of ~ 10%. This is a very great success for an organic system, but not good enough yet for mass commercialization which needs ~18%. Polymers can be made plastic, and even woven into garments, and made biocompatible, and this creates a wide range of new applications; for example, in biomedicine,"tattoo electronics" are already in use now. Building electronic circuits in the body and brain, self-powering these devices with PVC, is a great challenge, which is being pursued vigorously with recent applications to making wireless *wifi* brain to spine communication (Capogrosso et al. 2016), in order to cure paralysis and maybe blindness.

In the next chapter, we shall focus on another very exciting topic which is the harvesting of heat, either directly from the sun's infrared rays or from hot bodies created in, for example, "motors," friction, nuclear plants, or even geothermal processes. The heat ray part of the light spectrum is in the wavelength range longer than 5 μm; see Figure 1. Though one could in principle use semiconductors with very low bandgaps, the problem is that such devices would have a huge dark current for a given load which would swamp the photocurrent and deform the ideal square-shaped IV curve into a triangle with smaller fill factor.

Problems

1. Using the Hecht formula Eq. 11.1, calculate the carrier collection efficiency η given that the width of the device is 5 μm, the drift velocity is 10^5 cm/s, and the recombination time is 1 ns.
2. In your own words, explain how the multijunction cell illustrated in Fig. 11.5 a works. How do the two absorbing junctions cooperate to optimize the collection of light over a wider spectrum?
3. Define the power collection efficiency Λ in terms of its components and explain why a square-shaped IV curve is better than a triangular one.
4. If the carrier collection efficiency is 1, what is the single most important factor which limits the solar cell performance in a single junction system?

References and Further Reading

Berthold T et al (2012) Photosynthesis. J Am Chem Soc 134:5563

Capogrosso M et al (2016) A brain–spine interface alleviating gait deficits after spinal cord injury in primates. Nature 539:284–288. https://doi.org/10.1038/nature20118

Credgington D, Durrant JR (2012) Insights from transient optoelectronic analyses on the open-circuit cells. J Phys Chem Lett 3:1465–1478

Delaunay PY, Nguyen BM, Hoffman D, Razeghi M (2008) High-performance focal plane array based on InAs/GaSb superlattices With a 10m cutoff wavelength. IEEE J Quantum Electron 44:462–467.

Hecht K (1932) Zum Mechanismus Des Lichtelektrischen Primarstromes in Isolierenden Kristallen. Z Phys A 77:235–245

Jordan P et al (2001) Nature 411:909

Movaghar B, Schirmacher W (1981) On the theory of hopping conductivity in disordered systems. J Phys C solid state Phys 14:859

Polman A, Knight M, Garnett EC, Ehrler B, Sinke WC (2016) Photovoltaic materials: present efficiencies and future challenges. Science 352(6283)

Rosencher E, de Vinter B (2005). Optoelectronics. Cambridge University Press. 0521771293.

Sariciftci NS, Smilowitz L, Heeger AJ, Wudl F (1992) Photoinduced electron transfer from a conducting polymer to Buckminsterfullerene. Science 258:1474–1476

Savoie B, Movaghar B, Marks T, Ratner M (2013) Simple analytic description of collection efficiency in organic photovoltaics. Phys Chem lett 4:704

Street RA, Schoendorf M (2010) Phys Rev B 81(205):307

Yamaguchi M et al (2005) Multi-junction III-V solar cells: current status and future potential. Solar Energy 79:78

Thermal and Photothermal Energy Harvesting

12

12.1 Introduction

Turning away from fossil fuel to green energy is one of the most important targets for the world and critical to the survival of our way of life this century. Wind and solar energy are being developed in almost every country on the globe. They provide some relief but at this time fossil fuel is still the dominant source of energy. One way forward is to try to minimize the loss and wastage of energy, specially where fossil fuel is inevitable. Generally speaking one can summarize present-day wastage cycles as unused heat which is an inevitable component of most power sources: automobile engines, power stations, and solar radiation, just to mention a few examples. The reader is referred to the widely available literature on green energy and global for detailed descriptions of this cycle. Here we shall, as solid-state physicists, concentrate on two examples: (1) heat harvesting from hot surfaces or hot sources via the thermoelectric effect and (2) harvesting the sun's heat rays, i.e., the energy of the longer-wavelength part of the spectrum which is not routinely collected by commercial solar cells (Fig. 12.1).

This we call thermophotovoltaics as opposed to normal photovoltaics. Both these topics are very much part of solid-state discipline and involve the most modern nanotechnological considerations, with materials and material growth and structurization forming the key ingredients.

12.1.1 Power Generation

Approximately 90% of the world's electricity is generated by heat energy, typically operating at 30–40% efficiency, losing roughly 15 terawatts of power in the form of heat to the environment. Thermoelectric devices could convert some of this waste heat into useful electricity. Thermoelectric efficiency depends on the *figure of merit*, ZT. This will be discussed in a later section. There is no theoretical upper limit to ZT,

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_12

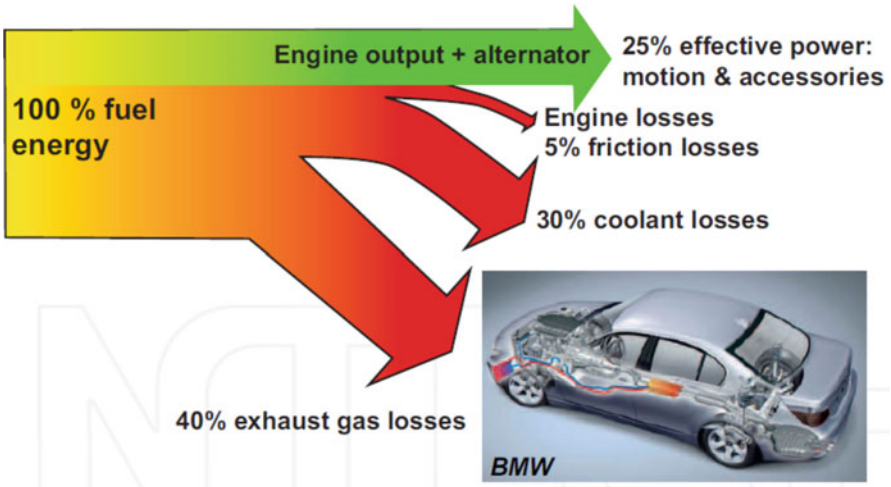


Fig. 12.1 A schematic diagram of how the energy from a combustion engine in a car is distributed Twenty-five percent of the energy produces motion and through the alternator generates electricity to power accessories including the electrics, the air conditioning, and the hifi system. Seventy-five percent of the energy from the fuel is lost mostly through friction and heat; 40% of the fuel energy disappears though the exhaust system; hence, there is interest in using thermoelectrics to harvest some of the waste energy (From Douglas Paul Book chapter http:dx.doi.org/10.5772/57092)

and as ZT approaches infinity, the thermoelectric efficiency approaches the Carnot limit. However, no known thermoelectrics have a ZT > 3. As of 2010, thermoelectric generators serve application niches where efficiency and cost are less important than reliability, light weight, and small size.

Internal combustion engines capture 20–25% of the energy released during fuel combustion. Increasing the conversion rate can increase mileage and provide more electricity for onboard controls and creature comforts (stability controls, telematics, navigation systems, electronic braking, etc.). It may be possible to shift energy draw from the engine (in certain cases) to the electrical load in the car, e.g., electrical power steering or electrical coolant pump operation.

12.1.2 The Thermoelectric Effect

Consider a piece of material, metal, semiconductor, glass, alloy, etc., and keep one end at a high-temperature T_h and the other end at a lower-temperature T_c. If a conducting wire is connected across the hot and cold ends, then a voltage will develop across the terminal, and power can be generated (Figs. 12.2 and 12.3).

The voltage or current generated by temperature differences is called the "thermoelectric voltage and due to the thermoelectric effect." The power extracted is called "thermoelectric power." Before proceeding to analyze the amount and efficiency of heat extraction, and the materials to be used, we need to recall some basic

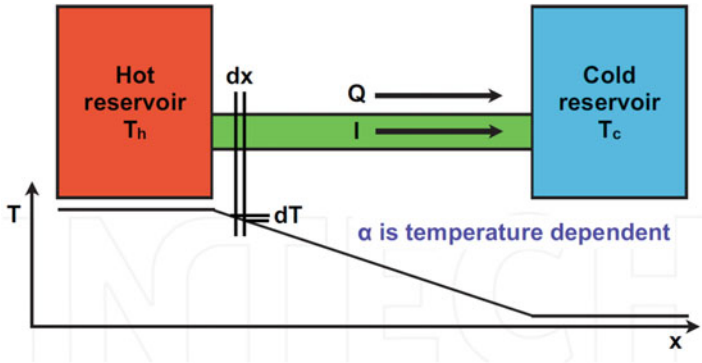


Fig. 12.2 In the above diagram, an electrical current I is generated as well as a heat current Q . As a result of this geometry, a temperature gradient $T(x)$ develops along the specimen so that the temperature difference depends on the position x along the specimen wire

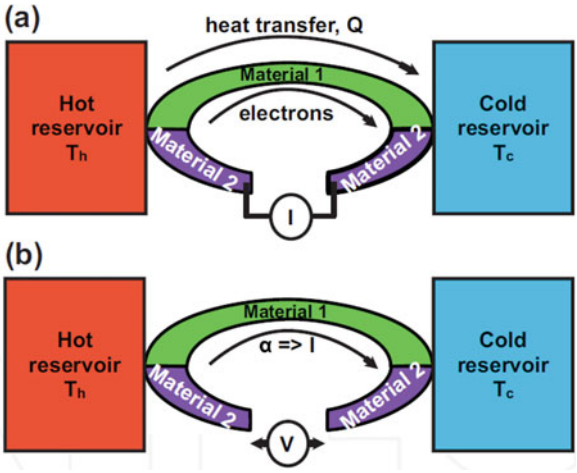


Fig. 12.3 (**a**) The thermocouple system between two heat reservoirs required to demonstrate the Peltier effect. (**b**) The thermocouple system between two heat reservoirs required to demonstrate the Seebeck effect (From Douglas Paul, Book chapter, http:dx.doi.org/10.5772/57092)

principles of transport theory. Let us start with heat transport. In the illustration shown, heat flows from hot to the cold end.

The amount of heat flowing Q is given by:

$$Q = -A\kappa \nabla T \tag{12.1}$$

where κ is the thermal conductivity in units of watts/Km and A the area with ∇T, denoting the temperature gradient. Along a wire of length L, the expression for the heat transferred Q is:

$$Q = -A\kappa\{T_h - T_c\}/L \tag{12.2}$$

The thermal conduction efficiency plays a critical role in thermal harvesting; we shall see how later. Let us now define the electrical conductance G:

$$G = A\left\{\int dE\left(-\frac{\delta f}{\delta E}\right)\rho(E)\sigma(E)\right\}/L \tag{12.3}$$

$$I = G\delta V \tag{12.4}$$

where $\sigma(E)$ is the energy-dependent electrical conductivity, $\rho(E)$ the density of states, and δV the voltage drop. The energy-dependent conductivity σ for bulk materials can be written (q is the electrical charge):

$$\sigma(E) = q^2 < v^2(E) > \tau(E) \tag{12.5}$$

with $<v^2(E)>$ denoting the mean squared velocity at energy E and $\tau(E)$ the effective carrier relaxation time. To recover the Drude theory at $T = 0$, one has to neglect all energy dependences and put $m^* < V^2 > \sim 2E$ the carrier energy, and then put $\rho E \sim n$ the carrier density, and replace the Fermi function derivative with a delta function.

In the limit of diffusive or hopping transport, it is convenient to introduce the carrier diffusivity $D(E)$ such that now:

$$D(E) =< v^2(E) > \tau(E) \tag{12.6a}$$

In the limit of quantum transport on the other hand, in short low-dimensional systems, it is more convenient to work with the Landauer transport formalism (see Razeghi Fundamentals Chap. 16). Here one works with the transmission coefficient $T(E)$ so that:

$$I = Aq^2\left\{\int dE\left(-\frac{\delta f}{\delta E}\right)\rho(E)T(E)V(E)\right)\}V \tag{12.6b}$$

$T(E)$ is calculated by considering the transmitted and reflected waves incident on the material; $V(E)$ is the velocity at energy E $(2E/m*)^{1/2}$, and V the external bias A is the area .

Finally the first principle expression for conductivity which can be reduced to Eq. (12.5) in the limit of bulk transport and scattering is the Kubo-Greenwood formula [Madelung, Solid-state physics] with the energy-dependent diffusivity given by:

$$D_\alpha = \hbar\sum_\beta \delta(\varepsilon_\alpha - \varepsilon_\beta)|\langle\alpha|v_x|\beta\rangle|^2 \tag{12.6c}$$

where v_x is the velocity operator in x-direction (same for y and z) and the matrix element is taken between the exact eigenstate of the system. Correlation effects if any are partially taken care of in the computation of the matrix element and appear as self-energy terms in the Green functions.

12.1.3 The Thermoelectric Voltage

The heat generated electrical current I_T can be calculated by envisaging the following scenarios: carriers in the hot region are excited to higher energies; they will have higher average velocities and will move, and delocalize, toward the cold terminal in order to equilibrate the slab. Similarly the fewer excited colder carriers will also delocalize toward the hotter terminal. The two constitute opposing currents which would exactly cancel if the temperatures were the same. But this is not the case, so there will be a net current flowing from hot to cold given by:

$$I_T = \Delta V_t G = \left\{ \int dE(f(T + \Delta T, E) - f(T, E))\rho(E)\sigma(E) \right. \tag{12.7}$$

$$f(T + \Delta T, E) = f_0 + \Delta T \frac{\delta f}{\delta T} \tag{12.8}$$

$$\frac{\delta f}{\delta T} = \frac{\delta f}{\delta E} \left\{ \frac{(E - E_f)}{T} - \frac{\partial E_f}{\partial T} \right\} \tag{12.9}$$

The derivative term involving the Fermi energy is neglected unless we are dealing with a strongly correlated system. So it follows that the voltage drop generated by the temperature difference can be written:

$$q\Delta V_t = \frac{T_h - T_c}{T} \left\{ \int dE \left(-\frac{\delta f}{\delta E} \right) (E - E_f)\rho(E)\sigma(E) \right\} / G \tag{12.10}$$

The Seebeck coefficient α is defined by the ratio of the voltage generated to the temperature difference:

$$\alpha = \frac{\Delta V}{\Delta T} \tag{12.11}$$

We note that the magnitude of α depends on the product of the energy from the Fermi level times the density of states times the conductivity. It is in effect the effective energy or entropy (times T) transported by the carrier as it moves from hot to cold. The Peltier coefficient is defined by:

$$\Pi = \alpha T \tag{12.12}$$

The units for the Seebeck coefficient are V/K. The Seebeck coefficient is $1/q$ times the entropy (Q/T) transported with each electron charge. Hence the Peltier effect is just due to electrons transferring heat from one reservoir to the other.

Note that a symmetrical density of states about the Fermi level gives zero net energy transfer and is therefore to be avoided by the designers.

It is instructive to examine some typical cases.

12.2 Seebeck Coefficient of a Free Electron Gas

This is given by:

$$\alpha = \frac{k_B}{q}\frac{k_B T}{E_f} \tag{12.13}$$

which is the classical energy per carrier $k_B T$ reduced by the effective number of carriers participating in the transport process, and the classical value of k_B /q is 86 μV/K.

12.3 The Seebeck Coefficient of an Undoped Semiconductor

In the nondegenerate limit where carriers have to be excited above the bandgap E_g of the semiconductor, the Seebeck coefficient is:

$$\alpha = \frac{k_B}{q}\frac{k_B T}{E_f} \tag{12.14a}$$

The thermoelectric efficiency to be developed in Section 3.0 is given by the product ZT:

$$ZT = \frac{\alpha^2 \sigma}{\kappa} T \tag{12.14b}$$

where κ is the *thermal conductivity* (Figs. 12.4 and 12.5).

12.4 Doped Semiconductors

The doping of semiconductors gives one a method of making conducting wires such that the charge transporter can be electrons or holes. This is beautifully illustrated in the device of Fig. 12.6, where electrons are the transporter of charge in the n-wire, but holes are the transporters in the p-wire. In this way it is possible to harvest heat into electricity or alternatively to transfer heat from the cold side to the heat sink and cool the material.

12.5 Seebeck Coefficient and Conductivity of a Hopping Conductor, i.e., Amorphous Silicon

The hopping transport limit is of some interest and importance because so many disordered materials exist which are hopping conductors. Amorphous silicon is a famous hopping conductor with considerable commercial significance. Its

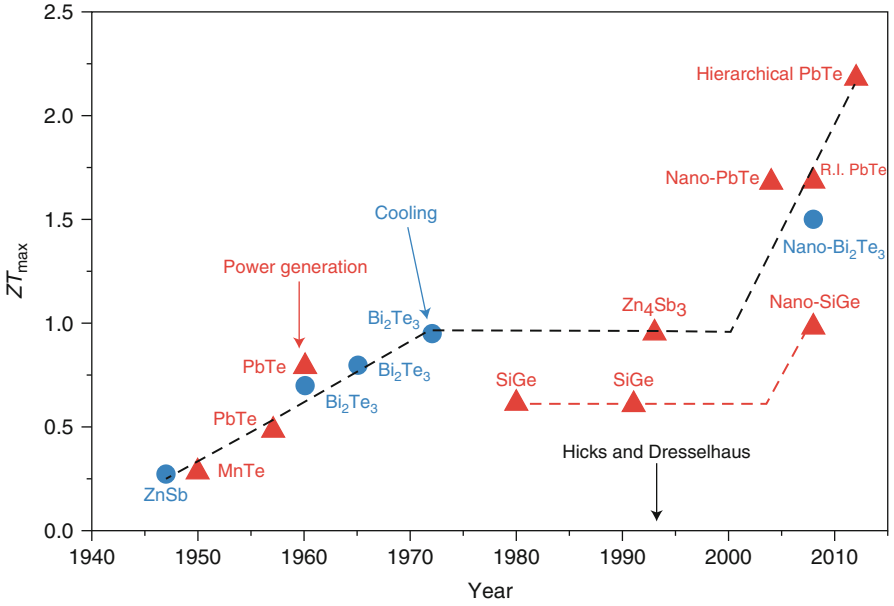


Fig. 12.4 Plot of achieved values of *ZT* versus year, defined in Section 3

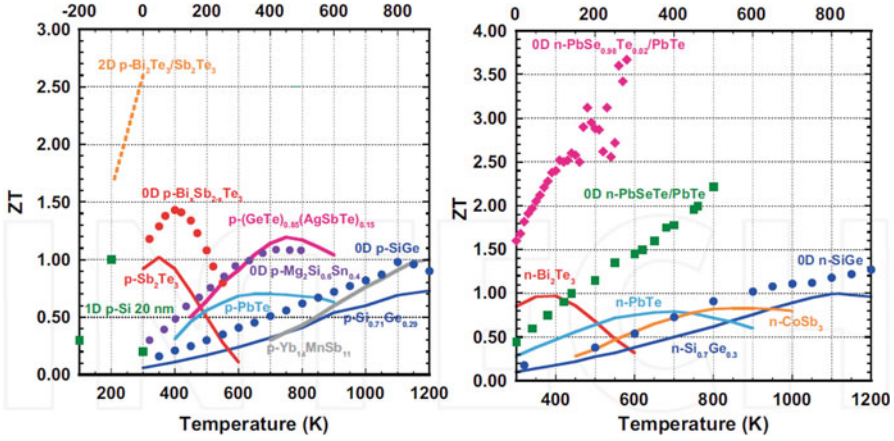


Fig. 12.5 Left: a comparison of *ZT* for p-type material as a function of temperature. Right: a comparison of *ZT* from n-type material as a function of temperature (See Douglas Paul book chapter (http:dx.doi.org/10.5772/57092 school of engineering Glasgow UK for original references))

conductivity exhibits the famous Mott $T^{1/4}$ law, and theoretician has been able to rigorously derive such laws and explain Mott's variable-range hopping (VRH) mechanism (see Mott and Davis Oxford Uni Press) (Fig. 12.7).

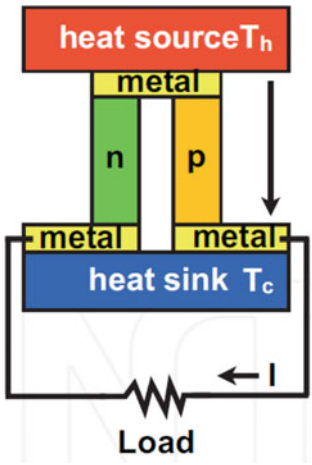


Fig. 12.6 Heat harvesting: electrons flow down the n-doped wire and are replenished by holes injected down the p-doped wire. The hot end helps to inject the carriers over the barriers

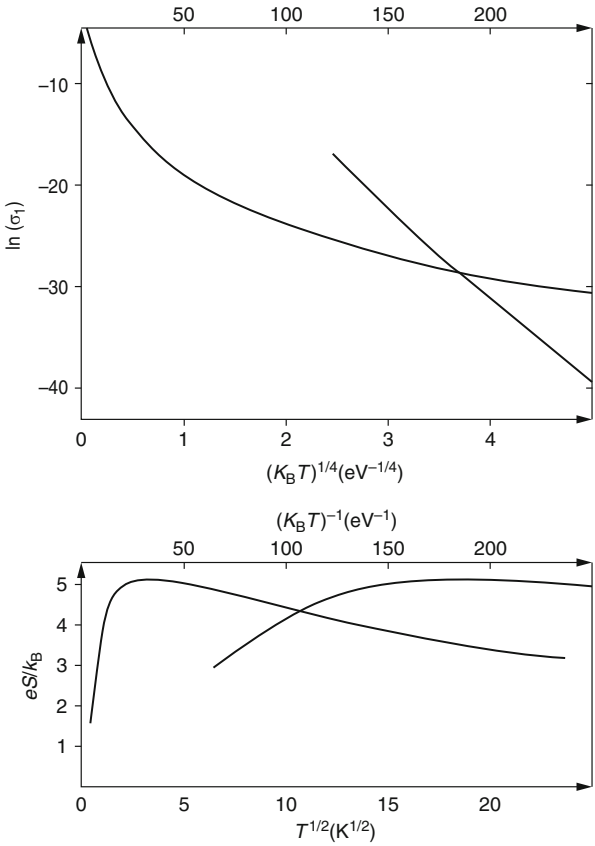


Fig. 12.7 Top curves: The log of the conductivity plotted against $1/T$ (left curve) and $T^{-1/4}$ (right curve). Lower: The hopping *thermopower* eS/k_b plotted against $1/T$ (left curve) and $T^{1/2}$ (right curve). The density of states used in the calculation is described in detail in Movaghar et al. and is linearly increasing in E from the Fermi energy (From Movaghar and Schirmacher)

12.6 Polaron Hopping

In the limit of strong electron-phonon interaction, the carriers will dig in and deform the lattice to lower the energy; the thermopower will now in general involve the transfer of both activation energy and vibrational energy. The latter depends on whether the sites are equivalent or not and on temperature. The details are beyond the scope of this article, and the reader is referred to the book by Emin or Boettger and Bryskin for a more complete discussion of this interesting subject.

In the following chapter, we consider the power that can be extracted from a thermoelectric circuit (Boettger and Bryskin (n.d.), Emin 1985).

12.6.1 Thermoelectric Efficiency

Let use imagine that we want work to be extracted from a heat engine; to analyze this we can go back to the basic principle of thermodynamics and consider the Carnot engine. Carnot showed a century ago that there is a maximum amount of work that can be extracted from a hot reservoir in a cycle of work such as a steam engine. Indeed this efficiency can be written:

$$\eta = 1 - \frac{T_c}{T_h} \tag{12.15}$$

where T_h is the temperature of the hot and T_c of the cold reservoir. This is related to the Kelvin statement of the second law of thermodynamics which states that no system operating in a closed cycle can convert all the heat absorbed from a heat reservoir into the same amount of work. Another way of saying the same thing is that no thermodynamic heat engine is 100% efficient.

Now one can think of the thermoelectric power generator as an engine and takes into account all gains and losses of energy in a complete cycle. Carriers are heated at the hot terminal, move to the cold terminal, and produce a voltage or current, but in the process they also dissipate joule heat because there is a current flowing. The inevitable heat flowing between the hot and cold reservoirs tends to equalize the temperatures. The heat reaching the cold terminal is not collected and therefore lost as far as the power generator is considered. When adding all gains and losses, one arrives at an efficiency equation which now looks as (Paul):

$$\eta = \left(1 - \frac{T_c}{T_h}\right) \frac{\sqrt{1 + zT} - 1}{\sqrt{1 + zT} + \frac{T_c}{T_h}} \tag{12.16}$$

where ZT is called the ZT factor and given by:

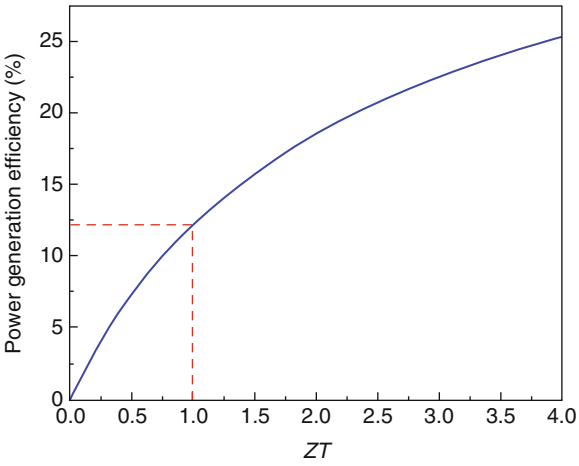


Fig. 12.8 Plot of power generation versus ZT as calculated from Eq. 12.16 using $T_h = 1273$ K and $T_c = 500$ K. The energy conversion efficiency for $ZT = 1$ is indicated by the dashed line

$$ZT = \frac{\alpha^2 \sigma}{\kappa} T \tag{12.17}$$

from which it follows that the thermoelectric efficiency is smaller than the Carnot efficiency and increases with the ZT factor of the material in question. In order to be of practical use the material must have a ZT of at least ~1. Figure 12.8 shows the connection between ZT and η .

Figure 12.8 data is from a dissertation presented by Xin Liang in Applied Physics, Harvard University, August 2013. Figure 12.8 shows calculation and plot of power generation efficiency with the figure of merit ZT; calculations were done using Eq. (12.16) with $T_h = 1273$ K and $T_c = 500$ K. The energy conversion efficiency corresponding to a ZT value of 1.0 is indicated on the plot.

The next observation is that in order to have a high ZT, a material must satisfy a number of conditions. First we note that the electrical conductivity and *Seebeck* coefficient α must be as high as possible. This must be achieved in combination with a thermal conductivity which is as low possible. Looking at the α for a semiconductor, one can conclude that this corresponds to a good value compared to a metal. This would be true except for the fact that a semiconductor with a high bandgap has a low conductivity which in turn lowers ZT and more than makes up for the gain in α. But as a rule one can see that materials with an asymmetric density of states around E_f which increases rapidly with a high value around 2 or 3 k_BT and covers delocalized conductive eigenstates are good for α.

The plot in Fig. 12.9 is very instructive and shows us how carrier density n, conductivity σ, and ZT scale. Surprisingly ZT seems to be best around a *Mott transition, i.e., a metal to insulator transition triggered by correlations*. But ZT also involves thermal conductivity κ in a significant way, so if we find a material with a good σ and α, for example, with Mott-like transition as shown, then we could proceed to lower its thermal conductivity by nanomaterial engineering. One way is to punch holes and make defects, cavities, and holes into the structure as shown in

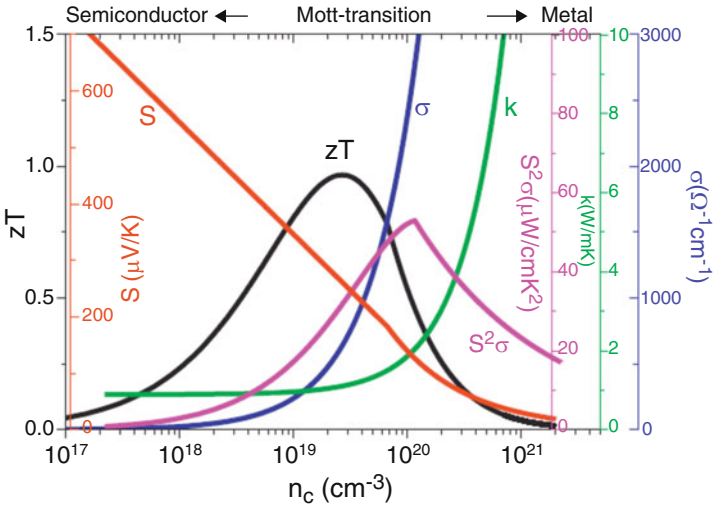


Fig. 12.9 A semi-schematic diagram of the thermoelectric properties using reasonable Seebeck coefficients S is here used to denote α the Seebeck coefficient (From "Ferroelectric thermoelectricity and Mott transition of ferroelectric oxides with high electronic conductivity," by Soonil Lee et al. Journal of the European Ceramic society vol.32, p3971 (2012))

Fig. 12.11 which illustrates how strongly the thermal conductivity can be made to vary using structural engineering. This kind of engineering will of course also change the conductivity, so in order to optimize *ZT*, the multifunctionality of the process has to be carefully considered. It is an exciting challenge for material engineering. In Sect. (12.7) we look at some actual values for *ZT* and see which ones are best. Let us now focus on one of the key properties, namely, the thermal conductivity.

12.6.2 Thermal Conductivity

The *thermal conductivity* of a material consists in general in two parts. It is the sum of the phonon or lattice contribution κ_{ph} and the electronic contribution κ_{e} (Keivan and Chen 2011).

The phonon thermal conductivity κ_{ph} can be calculated using the expression:

$$\kappa_{\text{ph}} = \frac{1}{3}\frac{C_v^{\text{ph}}}{V} v_p \Gamma \tag{12.18}$$

where C^{ph} is the phonon contribution to the heat capacity, V the volume, v_{p} a typical phonon velocity (velocity of sound), and Γ the phonon mean free path.

In a Bloch crystal at any temperature T, a more fundamental first principle expression is:

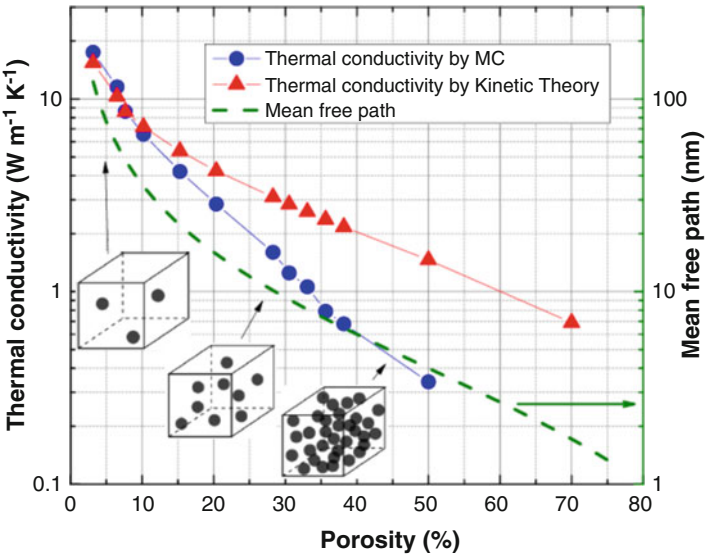


Fig. 12.10 Left axis: thermal conductivity variations of a mesoporous germanium thin film with a uniform pore diameter $d = 6$ nm as a function of the porosity; blue dots correspond to Monte Carlo simulations and red triangles correspond to the kinetic theory model. Right axis: nanoporous phonon mean free path as a function of the porosity for a uniform pore diameter $d = 6$ nm green dashed line (From "Thermal conductivity of meso-porous germanium" by M Isaiev et al. APL vol 105, 031912, (2014))

$$\kappa_{\text{ph}} = \frac{1}{3\Omega N_k} \sum_{k\lambda} v_{k\lambda}^2 \tau_{k\lambda} \hbar \omega_{k\lambda} \frac{\partial n_{k\lambda}}{\partial T} \tag{12.19}$$

where $\tau_{k\lambda}$ is the phonon relaxation time in a state $\mathbf{k}$, with λ denoting the mode index and Ω the unit cell volume, also is $n_{k\lambda}$ the Bose distribution function and ω_k the phonon frequency with v_k the corresponding velocity (Figs. 12.10, 12.11, 12.12, and 12.13).

Note: Just as the electrical conductivity, which was analyzed by Movaghar and Schirmacher, the thermal conductivity with disorder also increases with frequency until saturation sets in. Going up in frequency means sampling over smaller and smaller regions of the material and consequently more and more order. The heat diffusion theory presented below can offer an explanation of this behavior.

12.6.3 Thermal Conduction in the Diffusive Limit of Phonon Transport

In the incoherent propagation limit (short mean free path), heat transport can be described as an excitation diffusion process. We can apply the standard *temperature diffusion equation* given by:

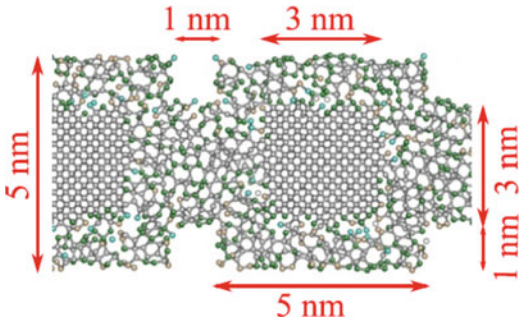


Fig. 12.11 The cross section of the modeled crystalline core/amorphous shell germanium nanoparticles with molecular dynamics is depicted. The characteristic lengths of the geometry are given. Gray atoms indicate four coordinated atoms, blue with one, yellow with two, and green with three (From "Thermal conductivity of meso-porous germanium" by M Isaiev et al. APL vol 105, 031912, (2014))

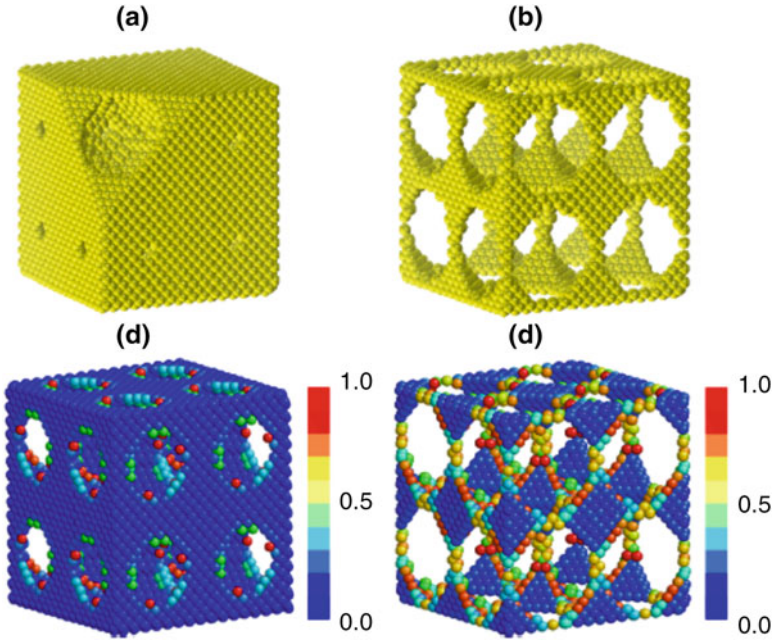


Fig. 12.12 Structure of nanoscale 3D Si PnCs . The period length of 3D PnCs is 8 units and the side length of simulation cell is 16 units. The periodic boundary condition is applied in simulation. The lattice constant is 0.543 nm of Si, and 1 unit represents 0.543 nm: (**a**) porosity is 50%, (**b**) has 90%, (**c**, **d**) normalized energy distribution on the PnC at 300 K with porosity of 70% and 90%, respectively (From: Extreme low thermal conductivity in nanoscale 3D Si *Phononic Crystal* with spherical Pores by Lina Yang et al. Nanoletters vol 14, 1734 (2014))

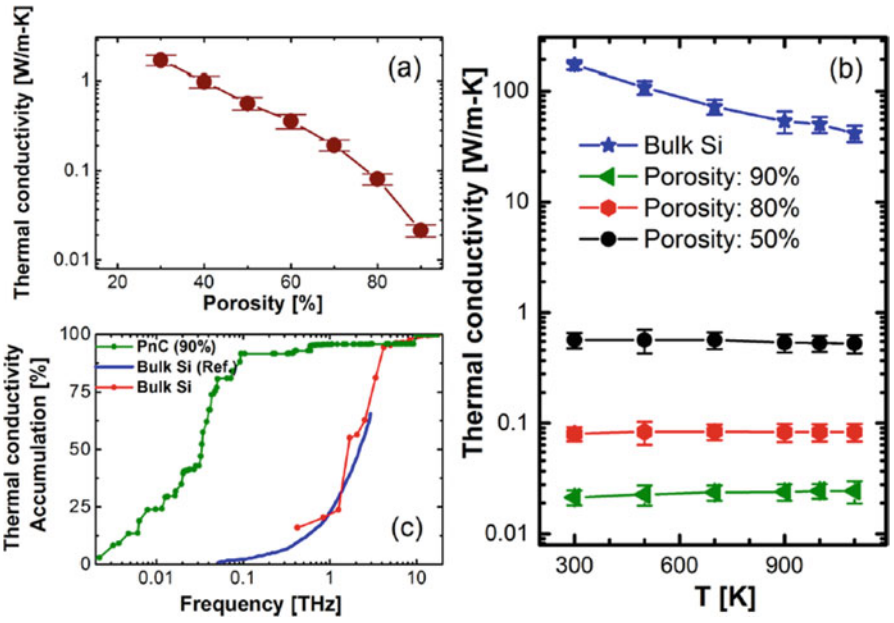


Fig. 12.13 (**a**) shows the strong connection between porosity and thermal conduction with (**b**) showing the temperature dependence, and the fact that when the disorder is strong, and the mean free path short, temperature or phonon-phonon scattering no longer matters. (**c**) Importantly the figure also shows the frequency-dependent thermal conductivity

$$\frac{\delta T\left(\vec{r}\right)}{\delta t} = \frac{\dot{Q}}{\rho C_\rho} + D_p \nabla^2 T - sT \tag{12.20}$$

where $T(r,t)$ is the local temperature at time t and where "s" is a uniform heat loss rate, dQ/dt is the rate of heating in a given region or point, ρ is density, C_ρ is specific heat capacity, and D_p is the thermal diffusion rate. In this limit it is the heat which is diffusing not the phonons. In the percolation or highly disordered limit, see Fig. 12.11; this equation can be discretized and turned into a hopping or random walk problem as done in Movaghar and Schirmacher for charge transport. The discretization can be done on a lattice for which the cell length is roughly the mean free path of the dominant phonons. We can Laplace transform Eq. (12.20) with "p" replacing "t" and solve this equation by Laplace transform on a lattice with constant D_p the (extra) temperature at "j" at time t given that the heat pulse was started at $r = 0$ at $t = 0$ is in Laplace space, with $\widehat{D_p}$ denoting the diffusion propagator matrix:

$$T_{0j}(p) = \left\langle j \left| \frac{1}{p + s - \widehat{D_p}/a^2} \right| 0 \right\rangle \left\langle 0 \left| \frac{\dot{Q}_i}{\rho C_\rho}(p) \right| 0 \right\rangle = \int_0^\infty dt e^{-pt} T_{oj}(t) \tag{12.21}$$

$$\frac{\dot{Q}_i}{\rho C_\rho}(p) \rightarrow g_0/p \tag{12.22}$$

For Eq. 12.22 it is assumed that the heating source has constant power *where* $\widehat{D}$ *"D- hat" is the diffusion matrix and "a," a cell-to-cell distance*. On a periodic lattice with Bloch dispersion ε_k, the diffusion propagator is:

$$G_{ij} = \left\langle j \left| \frac{1}{p + s - \widehat{D_p}\ /a^2} \right| i \right\rangle = G_{ij}(p) = \sum_k \frac{1}{p + s + \varepsilon_0 - \varepsilon_{\vec{k}}} \exp\left[i\ \vec{k}\ .\vec{R}_{ij}\right] \tag{12.23}$$

and represents the probability of a temperature shift "T" at the point R_j. It is in fact the (extra) temperature of the system at a point **R** in space given that the pulse was created at the point R_i at t = 0. Given a constant point power source generating heat at "0" with intensity g_0 in Hz, we have been using the standard diffusion solution in 3D (a = lattice spacing):

$$T(t,R) = g_0 \int_0^t d\tau \left(\frac{a^2}{4\pi D_p \tau}\right)^{3/2} \exp\left[-\frac{R^2}{4D_p\tau}\right] \exp[-s\tau] \tag{12.24}$$

The dispersion $\varepsilon(k)$ in Eq. 12.23 applies to a periodic lattice and represents the tight-binding-like energy versus k relation for the lattice in question; see Chaps. 2 and 5, Eq. 2.1 . For 3D cubic we recall ($t = D_p\ /a^2$):

$$\boldsymbol{\varepsilon}(\boldsymbol{k}) = 2|t|\left(\cos k_x a + \cos k_y b + \cos k_z c\right)$$

The case with disordered and spatially variable D_p can be solved like the corresponding hopping charge transport problem, in the same way. Here we use the effective phonon (heat) transfer rates from cell i to cell j, $D_{ij}/<a^2>$ rather than the charge hopping rates from site to site. The self-consistency relation involving the effective frequency-dependent heat diffusivity $D_p(\omega)$ for the "heat transfer rate" and the distribution function for the actual local transfer rates $D_{ij}/<a^2>$ will lead to an equation for the effective frequency-dependent phonon diffusion propagator $D_p(\omega)$ which contains the information on the thermal conductivity.

Given by:

$$D_p(\omega) = < \sum_j \frac{1}{1/D_{0j} + 1/\left(i\omega + D_p(\omega)\right)} > \tag{12.25a}$$

where $D_p(\omega)$ is the complex self-consistent average diffusivity, the average <>:

$$< Q >= \int_0^\infty dW_{ij} \chi\left(W_{ij}\right) Q \tag{12.25b}$$

goes over the distribution χ of transfer bonds W_{0j}, "j" is the sum index over the number of "transfer bonds" (heat hopping links) emanating from a given site "0", and $< a^2 >$ is the mean squared distance between the transfer cells. The calculation of D_p is completely analogous to the one for charge diffusion as given in ref Movaghar et al. The zero frequency thermal conductivity κ_{ph} is:

$$\kappa_{\text{ph}} = \frac{1}{3}\frac{C_p}{V} D_p(0) = \frac{1}{3}\frac{C_p}{V} v_p \Lambda_p \tag{12.26}$$

where v_p is the dominant phonon velocity (velocity of sound) and Λ_p the mean free path. The frequency diffusion rate $D_p(\omega)$ is complex, but the zero frequency $\omega = 0$ (long time) value is real. From the corresponding analysis of the ac hopping and percolation conduction, one can infer that with a suitable distribution function of phonon-hopping rates, Eq. (12.25) will be able to explain the behavior shown in Fig. 12.13c. At long times the diffusivity is lower because the phononic excitations have encountered the worse possible scenarios. At high frequencies the diffusivity reflects the well-connected and ordered domains where heat or phonon transport is highest.

In metallic materials we have conductivity contributions both from electrons and from phonons; here we show an example.

12.6.4 Phonon Contribution to Thermal Transport at Room T

$$\Lambda_p = 3.10^{-6}\text{cm}; v_p = 10^5\text{cm/s}; C_p/V = 25\text{J/K Mol}$$

$$\kappa_{\text{phonon}} = \frac{1}{3}\frac{C_p}{V} v_p \Lambda_p = 2.5 \ \text{W/cm/K} \tag{12.27}$$

12.6.5 Electron Contribution for a Metal at Room T ($C_{p,e}$ Is the Electronic Specific Heat)

$$\Lambda_e = 10^{-5}\text{cm}; v_e = 10^8\text{cm/s}; C_{p,e}/V = 0.5 \ \text{J/K Mol}$$

$$\kappa_e = \frac{1}{3}\frac{C_{p,e}}{V} v_e \Lambda_e = 250 \ \text{W/cm/K} \tag{12.28}$$

In materials with low electrical charge density, only the lattice or phonon thermal conductivity matters.

12.7 Summary: Typical Thermoelectric Generator

Using n- and p-doped semiconductors as charge/heat transporters, the common device structure for power generation is shown below (Figs. 12.14 and 12.15).

12.8 Application to Cooling

To cool a surface, we have to reverse the process and use electrical power to drag carriers from the side to be cooled to the heat sink; see Fig. 12.16 right. Best thermoelectric cooling at present works down to roughly 200 K which is impressive when taking into account the simplicity of the set up.

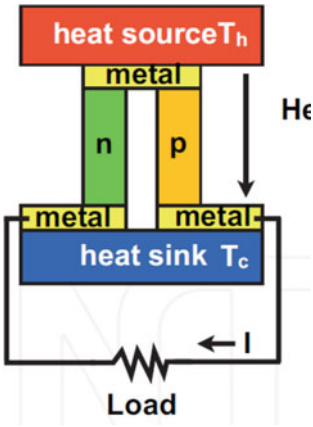


Fig. 12.14 Schematic energy diagram of a basic thermocouple unit made with doped semiconductors for harvesting heat

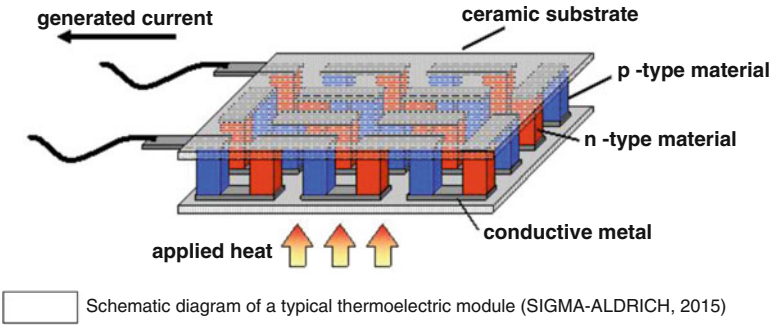


Fig. 12.15 Schematic diagram of a typical thermoelectric module (Sigma-Aldrich) (Sigma-Aldrich (2015) Materials for advanced thermoelectrics. Retrieved from http://www.sigmaaldrich.com/materials-science/metal-and-ceramic-science/thermoelectrics.html)

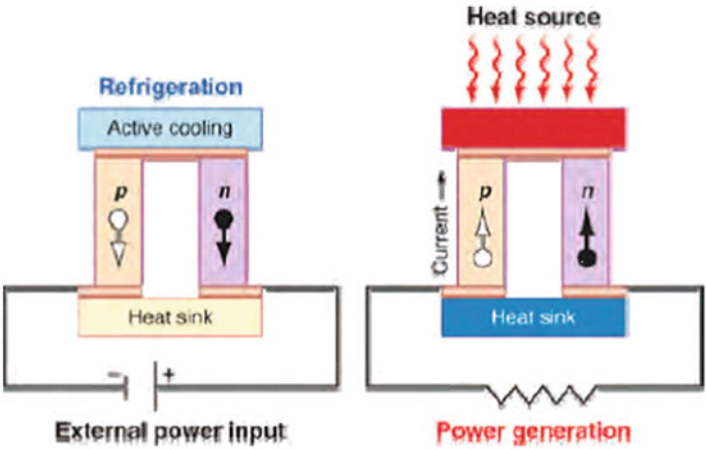


Figure 62. Thermoelectric refrigeration and power generation. A single thermoelectric couple is shown, configured for refrigeration (left) or power generation (right). The labels "p" and "n" refer to the sign of the majority charge carriers in each leg; "○" correspond to holes, and "●" correspond to electrons. The copper-colored regions depict electrical connections. Reproduced with permission from ref 650. Copyright 2002 American Association for Advancement of Science.

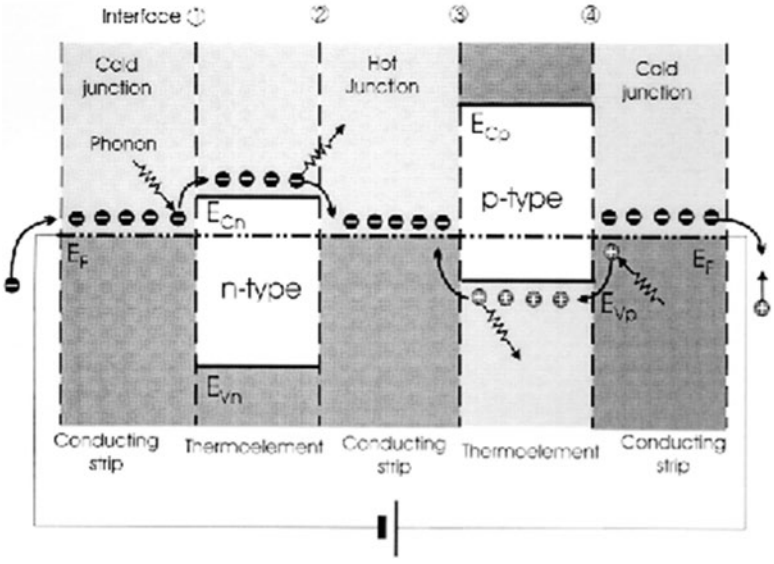


Fig. 12.16 The cold surface is cold because carriers absorb heat energy in order to transfer into the n-wire from the metal surface assisted by a bias. The electrons are replenished by cold electrons which are emitted by activated hole injection down the p-wire (From Paul). Both carriers in the top layer need heat in order to transfer along the circuit which is powered by a voltage source

12.9 Materials Old and New (Figs. 12.17 and 12.18)

From Douglas Paul's review, some of the original work comes from:

Dismukes JP et al (1964) Thermal electrical properties of heavily doped ge-si alloys up to 1300 K". J Appl Phys 35:2899
Venkatasubramanian R et al (2001) Thin film thermoelectrics devices with high room temperature figures of merit. Nature 413:597.
Boukai AI et al (2008) Silicon nanowires as efficient thermoelectric materials. Nature 451:168.

A more complete list of references is given in the review by D Paul.

12.9.1 Properties Which Make a Thermoelectric Material Efficient

The bismuth telluride (Bi_2Te_3)-type compounds are narrow gap layered semiconductors. The gap is ~0.2 eV with asymmetric density of states around E_f, making it ideal for carrier excitation at room temperature and high conductivity. Also, the van der Waals bonded layered structure is good for lowering the thermal conductivity. At all times one has to remember that electrical and thermal

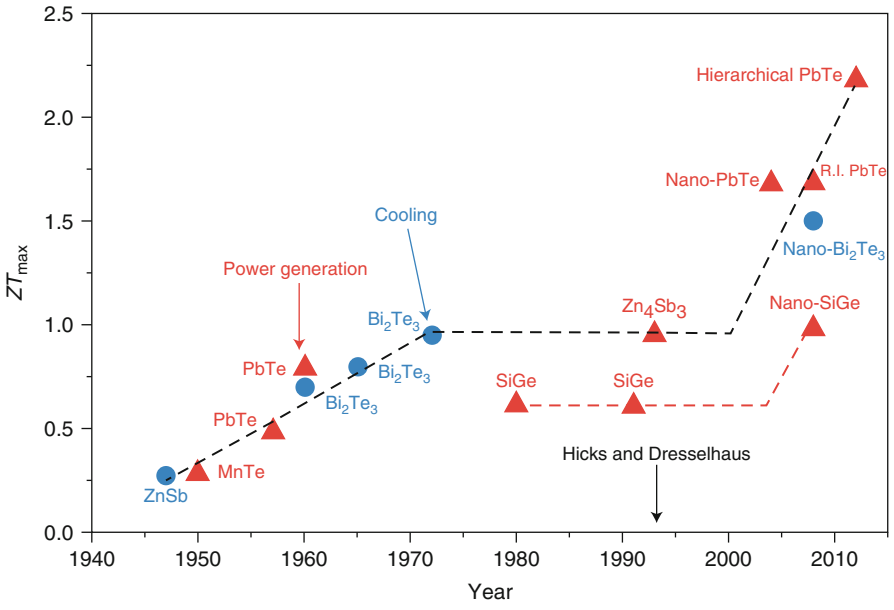


Fig. 12.17 From J P Heremans, M S Dresselhaus, L E Bell D T Morelli, "When thermoelectrics" reached the nanoscale" Nature Nanotechnology vol 8, p471, (2013)

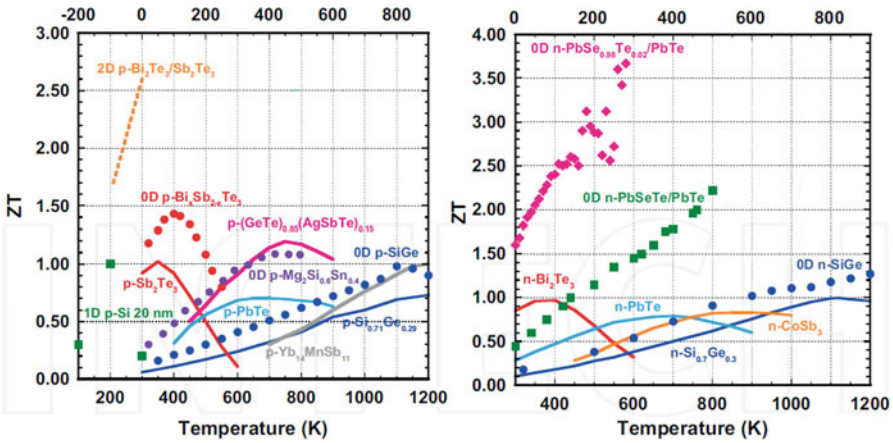


Fig. 12.18 Shows that one can achieve a few very good values of ZT with suitable material design. The inclusion of Te or Se seems to be particularly helpful. These alloys produce a high density of states just above the Fermi level which enhance α

conductivity of electrons in the relaxation time approximation are related to each other by the *Wiedemann-Franz law* which states that:

$$\kappa_e = \frac{\pi^2 k_b^2}{3q^2}\sigma \tag{12.29}$$

$$\kappa = \kappa_e + \kappa_p \tag{12.30}$$

This implies that the electronic part of the thermal conductivity is intimately linked to the electrical conductivity so that the two cannot be independently engineered. The total thermal conductivity is a sum of the electronic and phononic components, and the phononic component can be engineered independently. Also in semiconductors, usually $\kappa_e < \kappa_p$, lowering the phononic thermal conductivity is relatively straightforward because one can do that by lowering the crystal quality or dimensionality of the material.

12.9.2 Low-Dimensional Structures

Low-dimensional structures such as nanowires or quantum wells and superlattices can sometimes be used to optimize *ZT*. We saw, for example, how porosity helps to lower κ_p without necessary altering too much the electronic contributions. Superlattices can be made by depositing one material on top of the other, doping layers, forming minibands for electron transport, and indeed making phonon filters (Fig. 12.19).

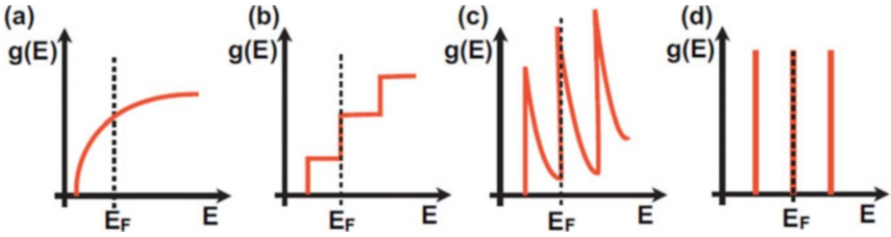


Fig. 12.19 The electron density of state as a function of energy for (**a**) 3D, (**b**) 2D, and (**c**) 1D semiconductor systems. The best position of the Fermi level for Seebeck coefficient is also shown (note the asymmetry) by the dashed line (see Dresselhaus et al. book for *ZT* "Recent trends in thermoelectric power conversion")

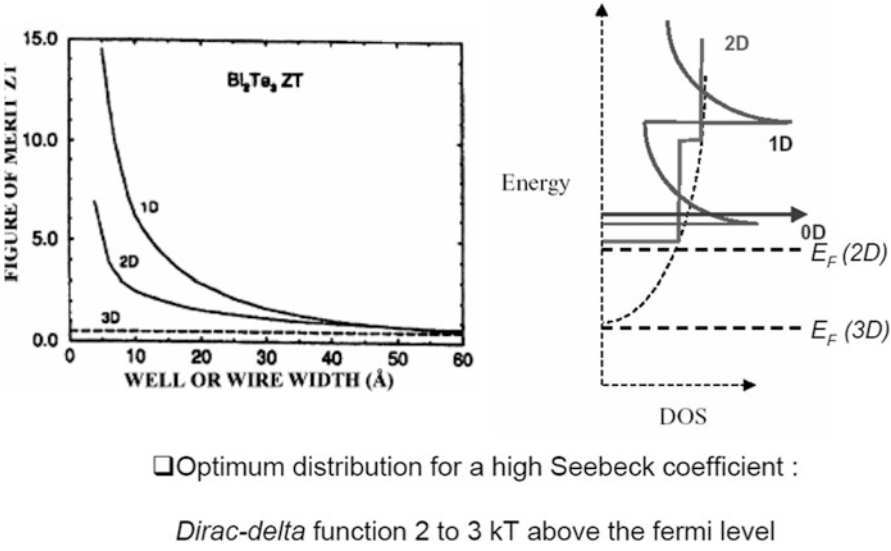


Fig. 12.20 Left figure shows theoretical estimates for the figure of merit as a function of well or wire width, from Dresselhaus; right figure shows the density of states and Fermi energy assumed (see Dresselhaus et al. book for *ZT* "Recent trends in thermoelectric power conversion")

A system may be designed and fabricated which optimizes *ZT* to the extent allowed by material properties. Phonon transport and electron transport in superlattices (SL) and superlattice wires are normally studied separately. The science of thermoelectric engineering of SL is relatively new. Samarelli et al. have shown that the expensive Te in BiTe materials can be replaced by SiGe modulation-doped superlattices with *ZT* reaching respectable values of *ZT*~0.15 (Fig. 12.20).

12.9.3 Advantages of Lower Dimensionality

These are associated with the lower thermal conductivity by enhanced phonon scattering and the strong asymmetry of the density of states about the Fermi level when strategically doped (Figs. 12.21 and 12.22).

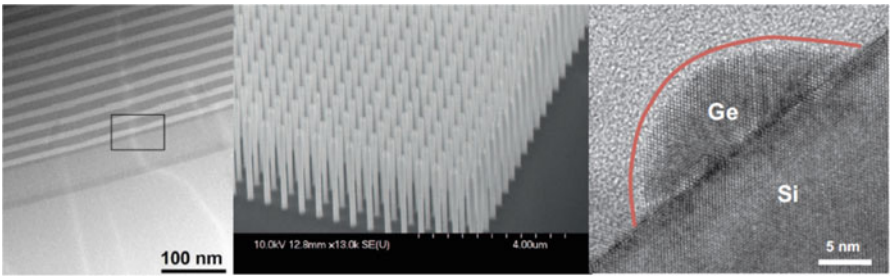


Fig. 12.21 Left: a cross-sectional TEM image of Ge quantum wells with $Si_{0.2}$ $Ge_{0.8}$ barriers forming a 2D thermoelectric system. Middle: a SEM image of etched 50-nm-wide nanowire of Ge/$Si_{0.2}$ $Ge_{0.8}$ material forming 1D thermoelectric systems. Right picture: a TEM image of a Ge quantum dot grown on a silicon substrate forming a 0D thermoelectric system for scattering phonons

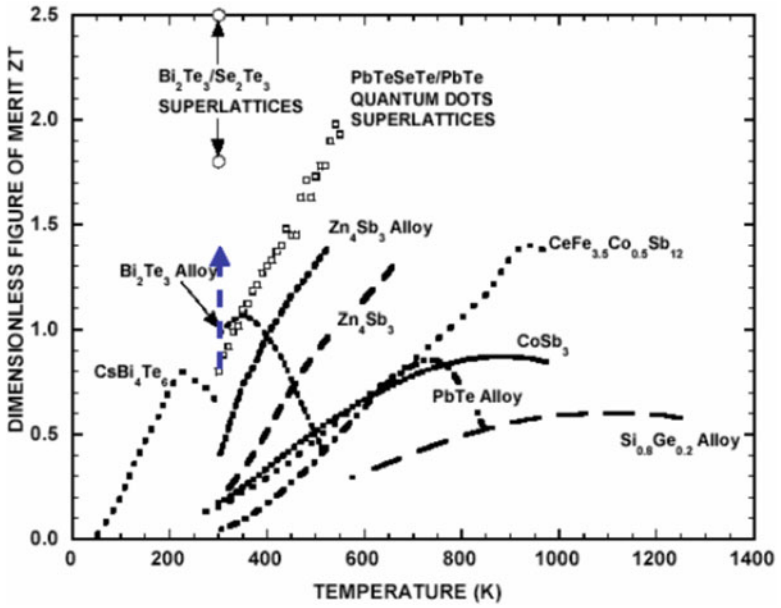


Fig. 12.22 The ZT value plotted for a number of alloys as a function of temperature ("Thermoelectric performance of metal semiconductor superlattice nanowires" from Sajid Kabeer. Online)

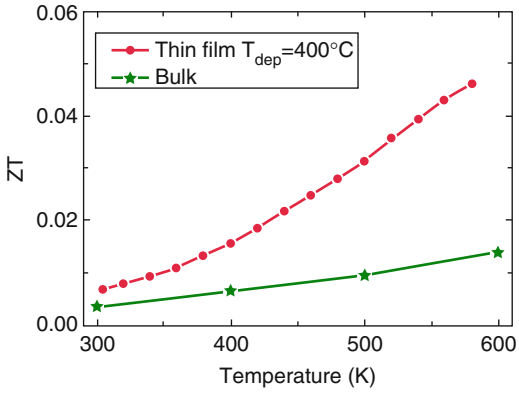


Fig. 12.23 Dimensionless figure of merit at elevated temperature using thermal conductivity at 300 K for AZO thin film deposited at 400C compared with previous material (see Shrikant Sani et al. Japanese Journal of applied Physics vol53 p 060306 (2014) "Enhanced thermoelectric performance of Al-doped ZnO films on amorphous substrates")

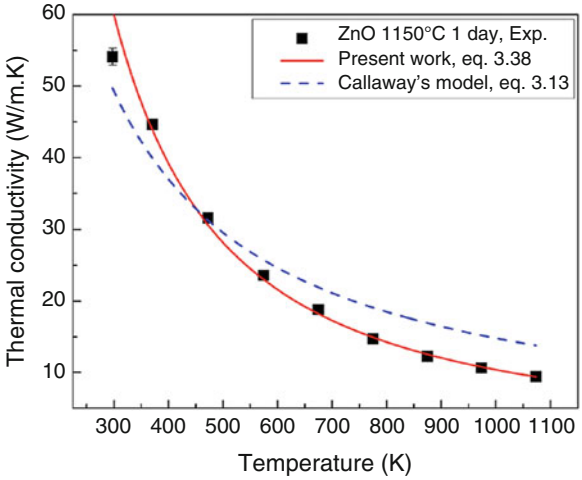


Fig. 12.24 A Samarelli et al. "the thermoelectric properties of Ge/SiGe modulation-doped superlattices" J Appl Phys Vol 113, 233704 (2013)

ZnO has a number of advantages (high bandgap, stability dopability) which can be exploited in thermoelectrics as well as shown by the Al-doped material in (Figs. 12.23 and 12.24)).

Thermal conductivity of mesoporous germanium by M Isaev et al. *APL, vol 105, p 031912 (2014)* (Fig. 12.25).

12.9.4 Summary

Low-dimensional structures such as nanowires can be beneficial because they can combine high electrical conductivity with low thermal conductivity using surface scattering of phonon heat carriers. Nanoparticle lattices are very versatile, and one has now learned how to generate assemblies with wide enough energy bands for high electrical conductivity, and soon one will learn how to concomitantly lower the

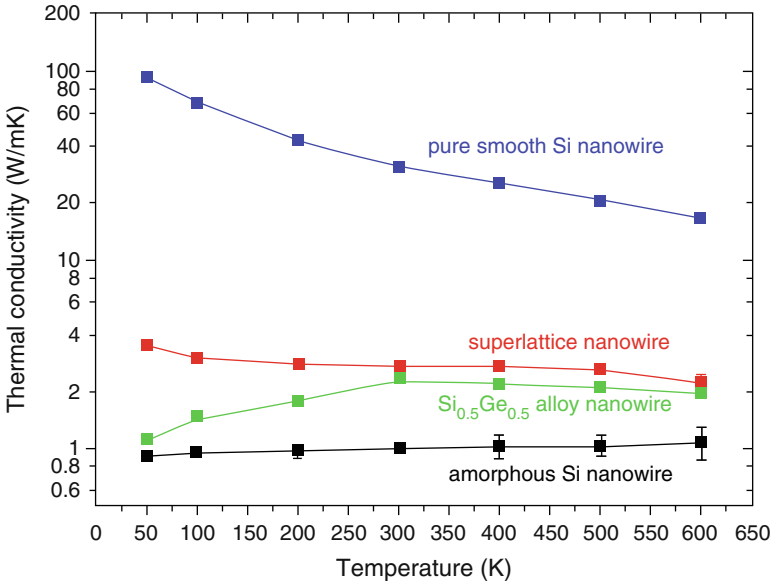


Fig. 12.25 Temperature dependence of the thermal conductivity of Si/Ge superlattice nanowires in comparison with pure smooth Si nanowires, $Si_{0.5}Ge_{0.5}$ alloy nanowires, and fully amorphous Si nanowires. All wires have the same cross-sectional area, 3.07nm × 3.07 nm, and the scan length of 278 nm from "Si/Ge superlattice nanowires with ultralow thermal conductivity" (Ming Hu and D Poulikakos Nanoletters, vol.1, p5487 (2012); see also Keivan Esfarjani and Gang Chen PRB vol 84 p 085204 (2011) Heat transport in silicon from first principles calculations")

thermal conduction. Metallic nanoparticles can be used as high absorbers of photons via surface plasmon excitations and hence used for nanoscale local heating. This can be integrated into nanotechnology to focus light, achieve strong absorption, and heating, and then hopefully high ZT value. There is a lot of scope in the nanoparticle material technology area, and there are some good prospects both for heat and light harvesting.

See Jong Soo Lee et al. (2011) Bandlike transport high electron mobility and high photoconductivity in all inorganic nanocrystal arrays. Nature Nanotechnology letters 6:348.

M P Bonechanscher et al. (2014) Long range orientation and atomic attachment of nanocrystals in 2D honeycomb superlattices. Science 344:1377.

References and Further Reading

Aldrich Sigma Aldrich (2015) Materials for advanced thermoelectrics. Retrieved from http://www.sigmaaldrich.com/materials-science/metal-and-ceramic-science/thermoelectrics.html

Boettger H, Bryskin V (n.d.) Hopping conduction in solids. VCH, Altenburg

Bonechanscher MP et al (2014) Long range orientation and atomic attachment of nanocrystals in 2D Honeycomb superlattices. Science 344:1377

Boukai AI et al (2008) Silicon nanowires as efficient thermoelectric materials. Nature 451:168
Dismukes JP et al (1964) Thermal electrical properties of heavily doped ge-si alloys up to 1300 K. J ApplPhys 35:2899
Emin D (1985) Polarons. Cambridge Univ press
Heremans JP, Dresselhaus MS, Bell LE, Morelli DT (2013) "When thermoelectrics" reached the nanoscale. Nat Nanotechnol 8:p471
Keivan E, Chen G (2011) Heat transport in silicon from first principles calculations. PRB 84:085204
Paul D. Thermoelectric harvesting, book chapter on line. http/dx.doi.org/105772/57092
Wan C et al (2015) Flexible n-type thermoelectric materials by organic intercalation of layered transition metal dichalcogenides with ZT of 0.28 at 338 K. Nat Mater 14:622

Photo-thermovoltaics

13

The other and more common strategy for harvesting energy from heat and light is to use PVC devices (see the previous chapter on light harvesting). The PVC devices are very well documented and constitute a mature technology that utilizes solar cells which the reader can access in the literature and books and buy in shops. The problem with current PVC technology is that it is mostly geared to the harvesting of shorter-wavelength region of the sun's spectrum; see Fig. 12.32. Silicon is currently still the typical and best material class; the lowest-energy photons collected have energies around 1 eV or 1.2 μm (Figs. 13.1 and 13.2, 13.3).

C Ferrari et al. "Overview and Status of Thermophotovoltaic Systems" Energy Procedia Vol 45, p160, (2014) 68th Conference of the Italian Machines Engineering Association ATI 2013

13.1 Photothermal Harvesting Using Photonic Crystal Conversion of Blackbody Heat into High-Energy Photons

The objective and design are illustrated in the diagram in Fig. 13.3. The idea is to harvest the long-wavelength light emitted by a hot body, not just the shorter wavelengths. Here, one designs an absorber of, for example, sunlight, the absorber gets hot, and now this absorber will also act as an emitter. Normally, it would emit in the blackbody spectrum. The point is to modify it in such a way so that its light emission now covers a smaller wavelength range than the whole blackbody spectrum and shifted to higher energies. In this way, the longer wavelengths can ideally be reemitted in a shorter wavelength range, and one can then proceed to harvest the heat rays emitted using conventional highly efficient semiconducting photovoltaic

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_13

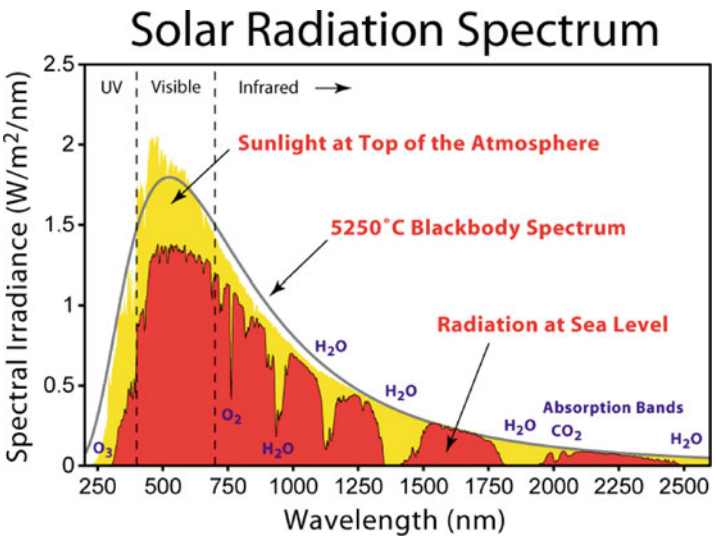


Fig. 13.1 Illustration of the sun spectrum

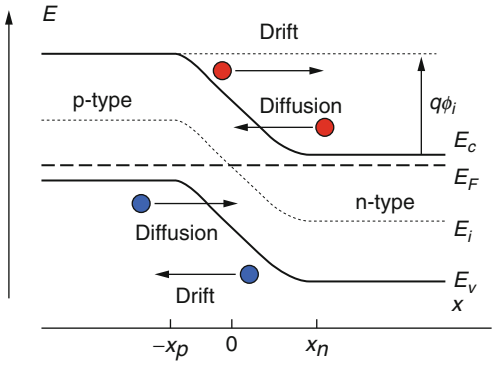


Fig. 13.2 p-n junction from Razeghi *Technology of Quantum Devices*

cells such as silicon cells. These cells operate above the bandgap of Si and thus cover the higher part of the bb spectrum.

In the proposed design shown in Fig. 13.5, the absorber and emitter consist of tungsten photonic crystals whose properties can be tailored to provide broadband absorption of light over the entire solar spectrum. The emission spectrum can be adjusted to match, at least in part, the absorption characteristics of a silicon photovoltaic cell. The absorber and emitter are integrated within the intermediate wavelength range to optimize thermal transfer. The bb light emitted which is below the absorption band of the photocell is designed to reflect back into the emitting material again; in this way it gets recycled into heat and is not lost. In this way, the conversion of long-wavelength light to shorter-wavelength light is done by the "emitter" itself, by way of raising the temperature of the emitter. This conversion does not explicitly

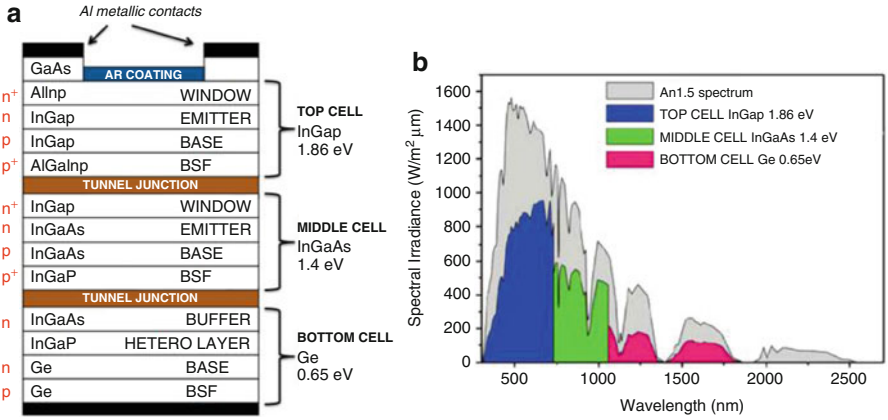


Fig. 13.3 Illustration of the action of a combination cell designed to pick up a wider range of wavelengths. The efficiency of such cells has reached values of 40% or more. The problem however is the cost of production which makes large-scale commercialization difficult

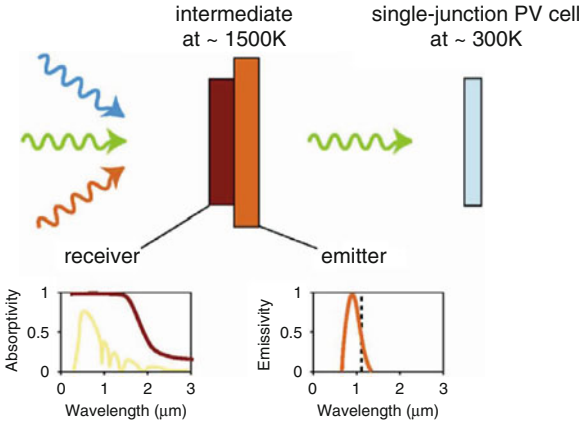


Fig. 13.4 Schematic of the proposed thermophotovoltaic cell. The receiver and emitter layers of the intermediate are shown together with the idealized absorption spectrum (including the solar spectrum in yellow) and emission spectrum (the dotted line indicate the PV cell bandgap energy level), respectively. The photonic crystal lens focuses the light, maximizes the absorption, and then reemits in a narrower wavelength region tuned to a conventional solar cell

involve exploiting any nonlinear processes, though these can indeed be involved as well (Fig. 13.4).

**The design and fabrication of photonic crystals transformers is a highly non-trivial task. The interference of light is used to narrow down the spectrum like Bragg reflection. This is in contrast to photonic conversion using the nonlinear optical effects, see the work of Fan* et al. *The net photon power emitted from a body at a given temperature T and emissivity σ_e is given by the Stefan-Boltzmann law*:

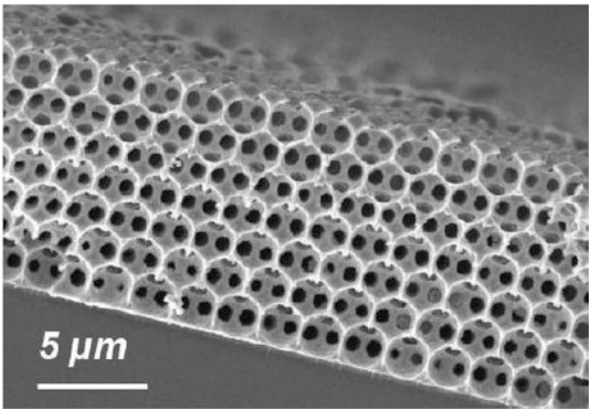


Fig. 13.5 Scanning electron microscope picture of a representative nickel photonic crystal fabricated by electrodeposition using a self-assembled polystyrene opal template

$$P_{\text{out}} = A\varepsilon\sigma_e\left(T^4 - T_0^4\right) = P_{\text{in}} \tag{13.1}$$

The trick is to keep the temperature high in the emitter using the recycled photons which are back-reflected into it. Photons can ideally only leak out through the allowed window and then into the photovoltaic device. The conversion to shorter wavelength is done by using the *entropic pressure* which makes a hotter body emit in a different (wider and shorter) spectrum with photons occupying higher energy modes with higher probabilities. The extra temperature is acquired by absorbing reflected photons from the photonic crystal filter.

13.2 Dichalcogenides: From Monolayers to Nanotubes

Layered compounds transition metal dichalcogenides TMD exhibit quite a good figure of merit as can be seen from Fig. 13.6, but they will, because of the high crystal quality, have high thermal conductivities which is a disadvantage. The latter needs to be reduced to enhance the ZT further.

Flexible n-type thermoelectric materials can be made by organic intercalation of layered transition metal dichalcogenide TiS_2 (Chunlei Wan et al. Nature materials vol 14 p 622 (2015)). Flexible n-type thermoelectric materials by organic intercalation of layered transition metal dichalcogenides with *ZT* of 0.28 at 338 K have been recently discovered; see above reference. The injection of organic layers considerably reduces the thermal conductivity.

*Organic layers were externally injected into the inorganic layers and then stabilized by organic cations, providing n-type carriers for current and energy transport. An electrical conductivity of 790 S cm(−1) and a power factor of 0.45 mW m(−1) K(−2) was obtained for a hybrid superlattice of TiS2/[(hexylammonium)x(H2O)y(DMSO)z], with an in-plane lattice thermal conductivity of 0.12 ± 0.03 W m(−1) K(−1), which is two orders of magnitude smaller than the thermal conductivities of the single-layer and bulk TiS2. High power factor and

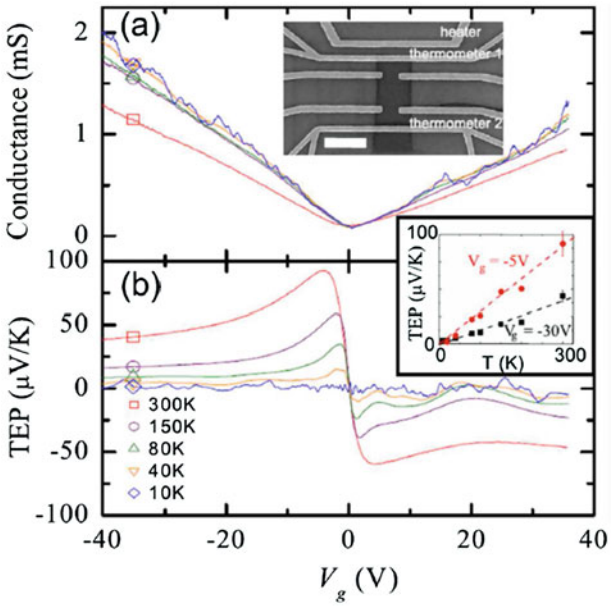


Fig. 13.6 (**a**) Electrical conductance G and (**b**) thermopower (α) TEP of a graphene sample as a function of back-gate voltage V_g for $T = 300$ K (square) 150 K (circle), 80 K (up triangle) 40 K (down triangle), and 10 K (diamond). Upper inset: SEM image of a typical device for thermoelectric measurements; scale bar is 2 μm. Lower inset: TEP values taken at $Vg = -30$ V (square) and -5 V (circle); dashed lines are linear fits to the data (From Y M Zuev et al. PRL vol 102 p 096807 (2009) Published on line Feb.5 (2014)) (From "Thermoelectric Properties of Transition Metal Dichalcogenides: From Monolayers to Nanotubes", Kai-Xuan Chen, Xiao-Ming Wang, Dong-Chuan Mo, and Shu-Shen Lyu. J. Phys. Chem. C, 2015, 119 (47), pp. 26706–26711)

low thermal conductivity contributed to a thermoelectric figure of merit, *ZT*, of 0.28 at 373 K, which might find application in wearable electronics.

13.3 Special Case: Graphene

Thermoelectric properties of graphene are described in the paper by Yong Xu et al. Condensed matter Science Feb 2015 published online. "Thermal and Thermoelectric properties of graphene."

The investigated thermoelectric properties on graphene suspended and on a substrate all demonstrate that high electrical conductivity is accompanied by high thermal conductivities as well, so that the *ZT* value is not very high. But individual properties are of great interest especially in view of the fact that a gate voltage can be used to control the carrier density, mobility, and conductivity and thus also the electronic thermal contribution. Researchers will eventually also find a way to also

control the lattice heat conduction by engineering defects and interfaces, superlattices, and heterojunctions in combination with TMDC, making graphene a uniquely versatile system.

13.4 Thermoelectric Mapping Graphene

Recently, researchers have managed to combine STM and thermal scanning to obtain atomic scale temperature images and have applied the technique to graphene. This is shown in Fig. (13.7); the temperature scan gives one yet another handle on the local properties of surfaces which can be combined with the STM and or AFM information to complete the picture. This is especially valuable for the study of systems which are highly correlated, such as magnetic layers or ferroelectrics, and materials which exhibit a metal insulator transition such as VO_2. One can in this way go some way toward disentangling the one-body from the many-body effects. But this type of work is very new, and a lot more needs to be done on measurement and modeling.

See "Atomic scale mapping of thermoelectric power on graphene: Role of defects and boundaries" Jewook Park et al., Nanoletters vol 13, p 3269 (2013).

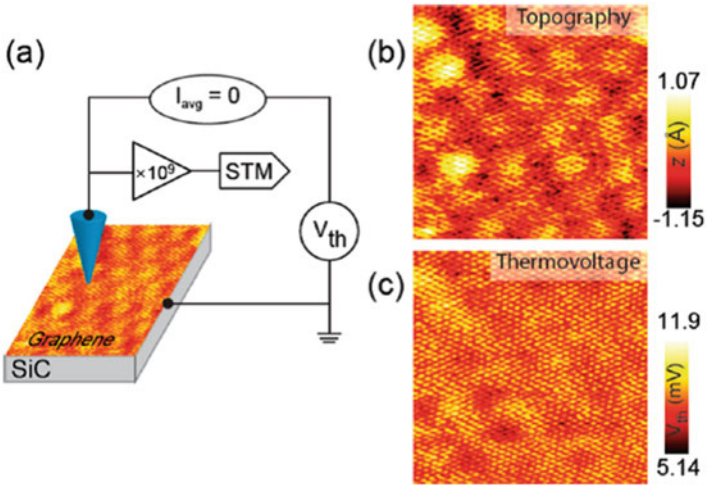


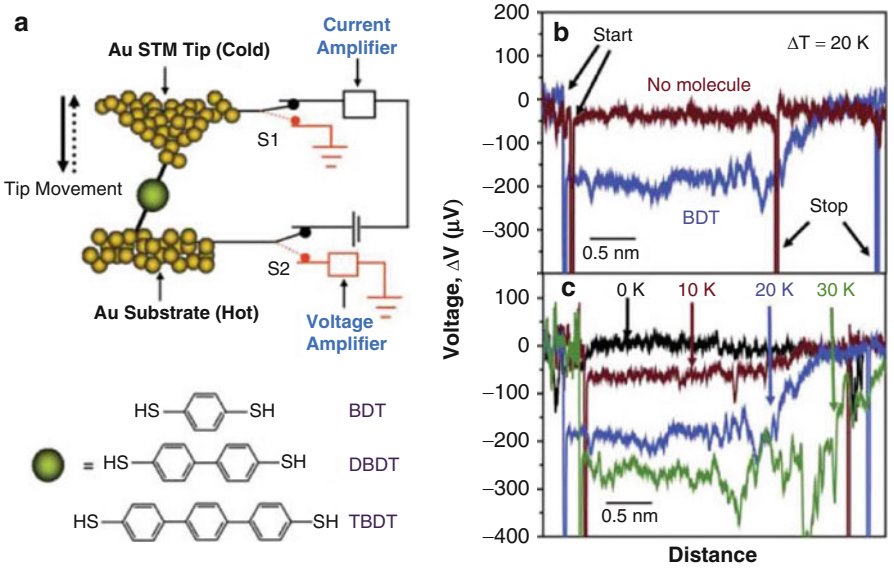
Fig. 13.7 Structure and thermovoltage measurement of graphene with an STM at the atomic resolution on epitaxial graphene on SiC. (**a**) Schematic diagram of measurement technique. Atomic-resolution images of topography (**b**) and thermovoltage (**c**) for the epitaxial graphene acquired simultaneously at 130 K (image size 7.5 nm × 7.5 nm). The temperature at STM tip is 298.5 K with an applied temperature difference delta $T = 168.5$ K. "Atomic scale mapping of thermoelectric power on graphene: Role of defects and boundaries" (Jewook Park et al., Nanoletters vol 13, p 3269 (2013))

13.5 Phononic Crystals

Another class of materials which is currently arousing great interest are the phononic crystals. Similar to photonic crystals, phononic crystals can be engineered with superlattices or other forms of topological complexity to form materials with phononic bandgaps. The field is relatively new and is very promising for application to green energy management and harvesting. It could well provide us with a highly desirable class of thermal diodes and selective frequency sound insulators or blockers. The reader is referred to the review by *Martin Maldovan in Nature vol 53 p 209* (2013).

13.6 Organic Materials: Single Molecule Junctions (Fig. 13.8)

Researchers have investigated structural, electrical, thermal, and thermopower properties of molecular junctions, down to single molecules; see Fig. 13.9. This is by now a vast field of research, with many interesting results (Reddy) published in the literature. Voltage generation is shown for a series of molecules in Fig. (13.9). Most work on molecule junctions at present targets, quite understandably *sensoric applications* for single molecule detection. As an engineering discipline, this subject is still in its infancy, but it is not too difficult to envisage the great potential that this type of approach has to offer. It must be put in conjunction and combined with nanoparticle surface plasmon technology, excitons, energy and charge transport



Figs. 13.8 and 13.9 From Pramod Reddy et al. "Thermoelectricity in Molecular Junctions" Science Vol 315 p 1568 (2007)

along DNA strands, muscle fiber, photosynthesis, nerve cells, tattoo electronics, animal hibernation, and information storage in biological materials. For transfer across a molecular bridge, see also D Segal et al. (J Phys Chem B vol 104 p3814 (2000)) and A B Butler Ricks et al. (JACS vol 132 p15427 (2010)).

13.7 Many-Electron Thermopower: The Effect of Electron Correlations

13.7.1 Kondo Systems

Up to now, we focused on materials which can be described by effective single particle physics. This includes the vast majority of useful semiconductors. The question is what happens when electron correlations get involved. This is an exciting subject in its own right. Though the Seebeck term can be large, useful thermoelectric performances are normally limited to low temperatures, because that is where correlations play a major role. Thus, this type of material technology is not currently pursued for large-scale energy harvesting and cooling. But in this category, we encounter a particularly interesting and topical class of materials, namely, solids, where narrow, strongly coulomb-correlated "d and f" electron bands are present which mix with broader s-like bands. The Hubbard energy U acts to oppose the filling of the atomic d or f shell leaving a net paramagnetic spin on the atoms. The localized spin strongly scatters the electrons in the s-bands, raising the energy of the electrons in such a way that the system prefers to screen this localized spin and form a singlet combination involving quasi-free "s"-like states and the more localized d-levels. The screened spin is then made invisible to the sea of other electrons roaming in the bands of the system, and this helps lower the total energy of the electrons. The localized singlet is called a "Kondo resonance" or Kondo bound state. This Kondo resonance produces a peak in the density of states for the excited spectrum which in turn leads to a high Seebeck coefficient. The problem is that the Kondo bound states are low temperature (<200 K) phenomena, and the *thermopower* is only enhanced typically at low temperatures, mostly below $T = 200$ K. Cooling and heat harvesting applications could indeed be envisaged at low T. For a comprehensive and up-to-date theoretical review, the reader is referred to the book by Zlatic. The Kondo systems $FeSb_2$ (see Fig) and $FeAs_2$ exhibit what can be termed a giant thermopower. The reason is that the s-d coupling or Kondo resonance generates a large density of states which peaks near the Fermi level. Looking at the formula (12.10), we note that this implies that excitation above E_f can lead to the transport of a large number of electrons which then gives rise to a large entropy per carrier. Kondo materials have a high and strongly temperature-dependent resistance (minimum structure) caused by the strong scattering of the nearly free electrons with the local bound state. It turns out that when magnetic impurities are used to dope nonmagnetic semiconductors giving rise to a Kondo resistivity, a similar peak in the density of states and similar enhanced value of the thermopower are recovered (Fig. 13.10).

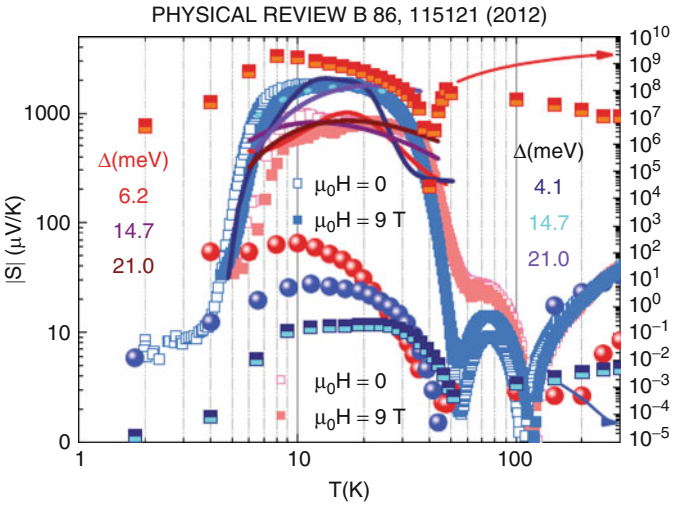


Fig. 13.10 $\alpha\,(T)$ marked on figure as S, for $FeSb_2$ crystal red and blue squares. Red and blue ball show fits to a correlated electron model. The ratio of individual and mobilities in a two-carrier model is shown by the red/orange and dark/light blue squares (see Qin Jiu et al.)

Qing Jie et al. "Electronic thermoelectric power factor and metal insulator transition in $FeSb_2$" PRB Vol 86 p115121 (2012), and references cited.

*Although small in comparison to α in $FeSb_2$, exceptionally large Seebeck coefficients are found in metals containing dilute magnitude impurities and in semiconductors containing resonant-level dopants and highly degenerate electronic bands. Among all these systems including $FeSb_2$, there is a shared commonality that plays a role in the large α, namely, large peak in the electronic density of states (DOS) near the Fermi level E_f. However, only in the former class of materials where magnetic interactions between localized and itinerant electrons take place are spins also suggested to play a role. Observation of this same effect, i.e., Kondo effect in semiconductors, has led to the term *Kondo insulators*. Examples of the anomalous peaks in α for these materials occur in FeSi, $Ce_3\,Pt_3\,Sb_4$, and $CeFe_4\,P_{12}$ reaching about 500, 350, and 800 μV/K, respectively (Fig. 13.11).

Tetsuro Saso "Thermoelectric Power and electronic structures of Kondo insulators" Physica B vol 328 p58 (2003)

"Correlated evolution of colossal thermoelectric effect and Kondo insulating behavior" by MK Fuccillo et al... APL Vol 1 062102 (2013)

JM Tomczak et al. "Thermopower of correlated semiconductors" Applications to $FeAs_2$ and $FeSb_2$ PRB Vol 82, p085104 (2010)

"Modern theory of Thermoelectrics" by Vejko Zlatic and Rene Monnier Oxford University press May 2014

Peijie Sun et al. Huge thermoelectric power factor: $FeSb_2$ versus $FeAs_2$ and $RuSb_2$, Applied Physics Express Vol 2 p 091102 (2009).

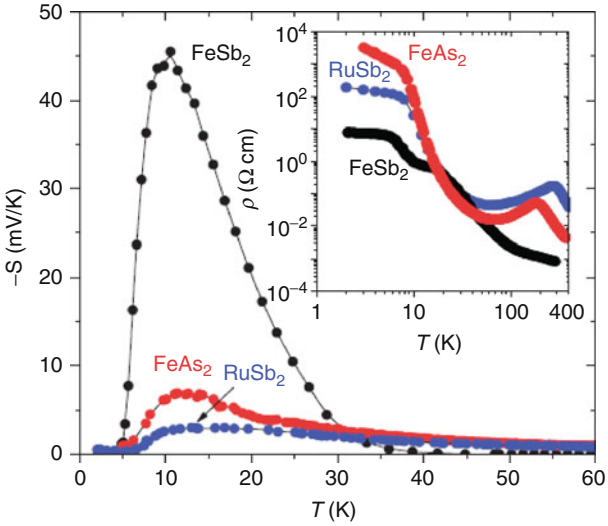


Fig. 13.11 Thermoelectric power S (α in our notation) for $FeSb_2$, $FeAs_2$, and $RuSb_2$ which all have a similar carrier concentration below 30 K; inset shows the electrical resistivity ρ as vs T. The resistance is thermally activated in certain T ranges. For $FeSb_2$ $E_g = 0.20$ eV. For $FeAs_2$ and $RuSb_2$ $E_g = 0.29$ eV (for original data, see Peijie Sun et al. Peijie Sun et al. Huge thermoelectric power factor: $FeSb_2$ versus $FeAs_2$ and $RuSb_2$, Applied Physics Express vol 2 p 091102 (2009))

13.8 Material with Metal Insulator MI Transitions, Example VO_2 Phase (Fig. 13.12)

In Fig. 13.13, we exhibit the evolution of the resistivity and Seebeck coefficients with temperature in VO_2.

A light beam applied to the material can switch the material from below to above the percolation threshold of conduction, and this gives a giant gain. Light can induce an MI transition in island networks. Understandably, this material is one of the best existing and useful bolometers because the MI transition is just above room temperature. For other examples, where collective effects give MI transitions, see also Fig. 12.9.

Takayaoshi Katase et al. "Thermopower analysis of metal insulator transition temperature modulation in vanadium dioxide films with lattice distortion."

*The conductivity can jump in the critical region when the gap decreases sharply with T.

*Movaghar-Schirmacher's method allows frequency-dependent thermal diffusion to be computed as well.

*Positive bias gate controlled metal insulator transition in ultrathin VO2 channels with TiO2 gate dielectrics by Yajima et al. DOI 10.1038 /ncomms10104

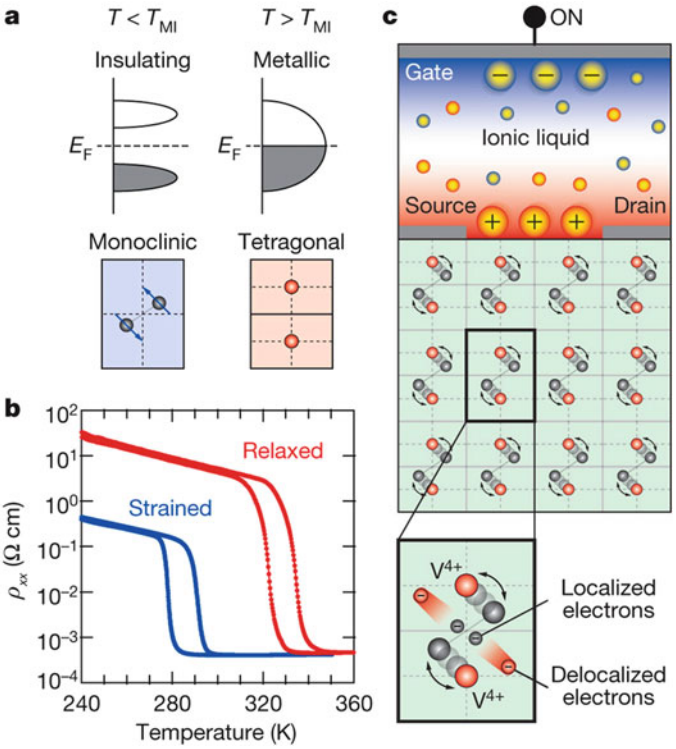


Fig. 13.12 Thermoelectric power S (α in our notation) for $FeSb_2$, $FeAs_2$, and $RuSb_2$ which all have a similar carrier concentration below 30 K; inset shows the electrical resistivity ρ as vs T. The resistance is thermally activated in certain T ranges. For $FeSb_2$ $E_g = 0.20$ eV. For $FeAs_2$ and $RuSb_2$ $E_g = 0.29$ eV (for original data, see Peijie Sun et al. Peijie Sun et al. Huge thermoelectric power factor: $FeSb_2$ versus $FeAs_2$ and $RuSb_2$, Applied Physics Express vol 2 p 091102 (2009))

13.9 Summary: Conclusion

Thermoelectric generators are already used in car exhausts and power stations and constitute now well-established technologies. But despite all the research, there are until now only a few materials which satisfy the efficiency criterion $ZT > 1$ criterion. Work is still in progress with new materials being designed and investigated every day using the most modern nanotechnology growth methods. The research program is exciting, and this applies in particular to the field of photothermal harvesting from blackbody sources.

In this chapter, we have shown the reader how to model thermoelectric efficiency, even, and in particular, in the presence of disorder. The chapter covered electric and thermal conductivity as well as the Seebeck effect.

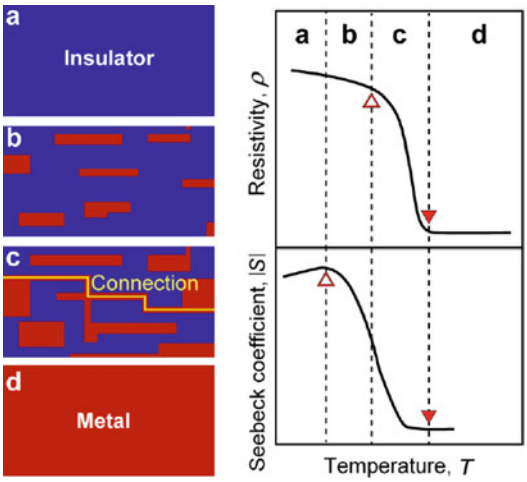


Fig. 13.13 Color online schematic illustration of metallic and insulating domain configuration at each temperature (**a**–**d**) around MI transition and corresponding changes in p-T and S-T Majority phase changes from Takayaoshi Katase et al. "Thermopower analysis of metal insulator transition temperature modulation in vanadium dioxide films with lattice distortion." Metallic domains are shown in red and grow as the T is increased until they form a connected percolation path. The thermopower and resistivity weigh in differently in this percolation transition, and the ensuing information is thus very useful. Thermal conductivity evolution will add yet another angle (Takayaoshi Katase et al., "Thermopower analysis of metal insulator transition temperature modulation in vanadium dioxide films with lattice distortion" transition in $FeSb_2$" PRB vol 86 p 115121 (2012))

For the materials physicists and theorists, the challenge is to identify, then model, and then fabricate the materials with a high Seebeck coefficient and high electrical conductivity yet low thermal conductivity. The effect of many-body interactions has also been considered. It turns out that Kondo insulators have high Seebeck coefficients, but their exploitation can at present only be envisaged at low temperatures. The physics of thermoelectric response in correlated materials such as ferromagnets is a subject of great interest in its own right because temperature gradients can produce ferromagnetic polarization and spin gradients without charge currents. Internal spin gates can be generated from regions of varying spin polarization, which seriously modify the internal carrier and phonon dynamics of the problem. This is especially true in ferromagnetic multilayers and nanowires. This subject is relatively new and as yet relatively unexplored. Applications are to be expected in the field of spintronics rather than energy harvesting (Nakano, Setsuro).

The fields of ordinary and giant magnetoresistance, magnetism and Hall effect, anomalous magnetoresistance, and spin Hall effect are by now vast and important fields of research and development in the category of spintronics. This book specializes in one-body quantum phenomena; spintronics involves collective effects and, like superconductivity, ferromagnetism, ferrimagnetism, and ferroelectrics, is beyond the scope of this book. The interested reader should consult some excellent

reviews, namely, by A Fert and P Gruenberg in *Nature Physics* vol 3 p 754 (2007); I Zitic, J Fabian, and S Das Sarma in Rev Mod phys vol 76, p323 (2004); and Thomas Dietl in *Nature Materials* vol 9 p 965 (2010) for the magnetically doped semiconductors. Albert Fert was awarded the Nobel Prize for his work on ferromagnetic interfaces and transport in 2007.

13.10 Discussion

In this chapter, we examined how electrical energy can be harvested from heat sources, for example, hot surfaces or substrates, and we did this in the context of solid-state electronics. The basic principle of thermoelectricity was explained, and it was shown how the thermally induced voltage could be calculated. As in a Carnot engine, some loss is inevitable, and the efficiency of a thermoelectric generator is lower than the Carnot efficiency. It was shown how the efficiency is calculated, and the figure of merit or ZT value was derived. The remainder of the chapter was devoted to studying particular examples. In particular, we discussed the effect of dimensionality, layered structures, molecular devices, and briefly also photothermal systems. It is still very difficult even now to generate useful systems with a $ZT > 1$, and the field is researched into with endeavor.

Problems

1. Calculate the thermoelectric efficiency ZT for a device for which the conductivity is 10^5 siemens/cm, the thermal conductivity is $\kappa = 5$ W/Km, and the Seebeck coefficient α is
   200 μV/k_b, $T = 300$ K.

$$ZT = \frac{\alpha^2 \sigma}{\kappa} T$$

   Write down the formula for the Seebeck coefficient of a free electron system and a wide bandgap semiconductor.

   Calculate it for crystalline Al metal and semic. AlAs at $T = 300$ K, extract parameters from Google.

2. (a) How does the spectrum of blackbody radiation scale with the temperature of a body? Make a typical sketch.
   (b)What is the photothermal effect? Explain how one can use the long-wavelength induced heating in a solid to harvest higher-energy photons in a conventional PVC.
3. Use your own ingenuity to design a thermoelectric material of high ZT. You can use any material and composition, drill holes, etc. Explain your choices. Plastic polymers are mechanically ideal, but what is the drawback in this context?

4. The Wiedemann-Franz law connects the thermal and electrical conductivity of a free electron gas in the Drude or nearly free electron gas approximation. The statement is (Eq. 6.38b):

$\kappa_{el} = \frac{\pi^2 k_b^2}{3q^2} T\sigma_{el}$ where κ_{el} is the thermal and σ_{el} the electrical conductivity.

From this law, it follows that one cannot expect a good simple metal to be a good *ZT* material because the thermal conductivity must be also high.

Calculate the thermal conductivity if the $\sigma_{el} = 10^4$ S/m at $T = 300$ K.

References and Further Reading

Ferrari FJ et al (2014) Overview and status of thermophotovoltaic systems. Energy Procedia 45:160 68 th Conference of the Italian Machines Engineering Association ATI 2013

Jeewok P et al (2013) Spin dependent Seebeck effect, thermal colossal magnetoresistance negative differential thermoelectric resistance in zigzag silicene nanoribbon heterojunction. Thermal conductivity of Mesoporous Germanium. Nanoletters 13:3269

Jie Q et al (2012) Electronic thermoelectric power factor and metal insulator transition in $FeSb_2$. PRB 86:115121 References cited

Martin M (2013) Sound and heat revolutions in phononics. Nature 53:209

Nakano M et al (2012) Collective bulk carrier delocalization driven by electrostatic surface charge accumulation. Nature 487:459–462. https://doi.org/10.1038/nature11296

Pramod R et al (2007) Thermoelectricity in molecular junctions. Science 315:1568 See also "Hybrid thermoelectrics".

Sun P et al (2009) Huge thermoelectric power factor: $FeSb_2$ versus $FeAs_2$ and $RuSb_2$. Applied Physics Express 2:091102

Tetsuro S (2003) Thermoelectric power and electronic structures of Kondo insulators. Physica B 328:58 "Correlated evolution of colossal thermoelectric effect and Kondo insulating behavior" by

Tomczak JM et al (2010) Thermopower of correlated semiconductors. Applications to $FeAs_2$ and $FeSb_2$. PRB 82:085104

Wan C et al (2015) Flexible n-type thermoelectric materials by organic intercalation of layered transition metal dichalcogenides with ZT of 0.28 at 338 K. Nat Mater 14:622

Xianliang L et al (2011) Taming the blackbody with infrared metamaterials as selective thermal emitters. PRL 107:045901

Zuev YM et al (2009) Atomic scale mapping of thermoelectric power on graphene. Role of defects and boundaries. PRL 102:096807 Published on line Feb 5 (2014)

Electron-Electron Interactions: Screening

14

14.1 Introduction

In this chapter we will investigate how the presence of other charges and dipoles influences the charge-charge interaction.

Consider, for example, a net charge introduced into a semiconductor, and consider how the electrons in the conduction band react to it. The net charge could be, for example, the charge of the ionized Si impurity in GaAs or P impurity in Si. One can ask what is the electric field that the carriers see? Do they see the full Coulomb field of the ionized impurity or a reduced field? We know from elementary electromagnetic theory that the presence of an insulating medium around a charge partially screens the charge and introduces the bound electron permittivity term ε_b in the Coulomb potential, for example. We also saw in Chap. 10 how one can derive the permittivity of bound charges ε_b in a medium. But what about the presence of free electrons? What happens to the net field in a metal where we have both bound and free charges? We saw how the free carriers change the total permittivity of a system and how this can be incorporated in the optical properties of solids, but we did not look at the consequences for carrier-carrier and carrier-charge interactions. The free carriers were given a classical Drude treatment which is adequate for optics, but we did not investigate how the medium affects the net interaction between the charges themselves. To answer this question, we have to start from a first principle point of view and give the problem a quantum mechanical treatment. Let us examine these questions from a fundamental point of view following closely the book by J. Ziman (see references). We start by applying a general potential $V(r, t)$ to a medium where:

$$V(\bar{r}, t) = V_0 e^{ik'\cdot\vec{r}} e^{i\omega t} e^{\alpha t} \tag{14.1}$$

We have allowed the field to grow slowly to its full value with a time constant α in order not to cause large deviations from equilibrium. Now we go back and use time-dependent perturbation theory as we did before for bound electrons in the previous

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_14

chapter and consider the first-order change in the wavefunction of a Bloch electron in a solid that this potential introduces (Ziman 1964):

$$\Psi_{\vec{k}}(\vec{r},t) = \left|\vec{k}\right\rangle + b_{\vec{k}+\vec{k'}}(t)\left|\vec{k}+\vec{k'}\right\rangle \tag{14.2}$$

whereas in Eq. (10.42), we have from perturbation theory:

$$b_{\vec{k}+\vec{k'}}(t) = \frac{\left\langle \vec{k}\left|V(\vec{k'},\vec{r},t)\right|\vec{k}+\vec{k'}\right\rangle}{E_{\vec{k}} - E_{\vec{k}+\vec{k'}} + \hbar\omega - i\hbar\alpha} \tag{14.3}$$

$$b_{\vec{k}+\vec{k'}}(t) = \frac{V_0 e^{i\omega t} e^{\alpha t}}{E_{\vec{k}} - E_{\vec{k}+\vec{k'}} + \hbar\omega - i\hbar\alpha} \tag{14.4}$$

The new wavefunction implies also a new charge distribution. Thus, we can compute by considering the deviation from the unperturbed distribution:

$$\delta\rho(\vec{r},t) = q\sum_{\vec{k}}\left\{\left|\Psi_{\vec{k}}(r,t)\right|^2 - 1\right\} \tag{14.5}$$

Substituting from Eq. (14.4) and keeping only terms in first order, we find:

$$\delta\rho(\vec{r},t) = q\sum_{\vec{k}}\left\{b_{\vec{k}+\vec{k'}}(t)e^{i\vec{k}.\vec{r}} + b^*_{\vec{k}+\vec{k'}}(t)e^{-i\vec{k'}.\vec{r}}\right\} \tag{14.6}$$

In practice, it is more convenient to work with a real perturbation, so let us write instead:

$$\delta V(\vec{r},t) = V_0 e^{i\vec{k'}.\vec{r}} e^{i\omega t} e^{\alpha t} + V_0^* e^{-i\vec{k'}.\vec{r}} e^{-i\omega t} e^{\alpha t} \tag{14.7}$$

Then it follows by substituting Eq. (14.7) into Eq. (14.3) and then Eq. (14.5) that:

$$\delta\rho = q\sum_{\vec{k}}\left\{\frac{V_0}{E(\vec{k}) - E(\vec{k}+\vec{k'}) + \hbar\omega - i\hbar\alpha} + \frac{V_0}{E(\vec{k}) - E(\vec{k}-\vec{k'}) - \hbar\omega + i\hbar\alpha}\right\}$$
$$\times e^{i\vec{k'}.\vec{r}} e^{i\omega t} + cc \tag{14.8}$$

The next step is to generalize this expression taking into account the fact that the initial states must be occupied, and the final states, to which the electrons are moved to by the perturbation, must be empty to find the charge density change:

$$\delta\rho = qV_0 \sum_{\vec{k}} \left\{ \frac{f_0\left(\vec{k}\right) - f_0\left(\vec{k} + \overline{k'}\right)}{E\left(\vec{k}\right) - E\left(\vec{k} + \vec{k'}\right) + \hbar\omega - i\hbar\alpha} \right\} e^{i\vec{k'}\cdot\vec{r}} e^{i\omega t} + cc \tag{14.9}$$

This is a new charge distribution caused by the application of the perturbation, so it also produces a new potential, which must be a solution of the Poisson equation:

$$\begin{aligned} \nabla^2(\delta\Phi) &= -4\pi q\delta\rho \\ \delta\Phi &= \Phi_0 e^{i\vec{k'}\cdot\vec{r}} e^{i\omega t} + cc \end{aligned} \tag{14.10}$$

where we have assumed that the potential has the same time and spatial variation as the perturbation. Now we substitute Eq. (14.10) and evaluate the ∇^2 operator to find:

$$\begin{aligned} -k'^2\Phi &= -4\pi q^2 V_0 \sum_{\vec{k}} \frac{f\left(\vec{k}\right) - f\left(\vec{k} + \vec{k'}\right)}{E\left(\vec{k}\right) - E\left(\vec{k} + \vec{k'}\right) + \hbar\omega - i\hbar\alpha} \\ \Phi_0 &= 4\pi q^2 \frac{V_0}{k'^2} \sum_{\vec{k}} \frac{f\left(\vec{k}\right) - f\left(\vec{k} + \vec{k'}\right)}{E\left(\vec{k}\right) - E\left(\vec{k} + \vec{k'}\right) + \hbar\omega - i\hbar\alpha} \end{aligned} \tag{14.11}$$

So we see that the perturbation has produced a reaction, a new internal potential. But this reaction is itself a similar perturbation, so the calculation should really be self-consistent and take this internal response into account right from the start. In other words, the total perturbation acting on the electrons is not just the external potential but also the internal response that the external one has generated. We now have the total perturbation:

$$\delta U\left(\vec{r}, t\right) = \delta V\left(\vec{r}, t\right) + \delta\Phi\left(\vec{r}, t\right) \tag{14.12}$$

And if we assume that the external potential has the form given by Eq. (14.7):

$$U = V + \frac{4\pi q^2}{k'^2} \sum_{\vec{k}} \left\{ \frac{f\left(\vec{k}\right) - f\left(\vec{k} + \vec{k'}\right)}{E\left(\vec{k}\right) - E\left(\vec{k} + \overline{k'}\right) + \hbar\omega - i\hbar\alpha} \right\} U \tag{14.13}$$

or in other words:

$$U = \frac{V}{\varepsilon\left(\vec{k'}, \omega\right)} \tag{14.14}$$

where:

$$\varepsilon\left(\vec{k'}, \omega\right) = 1 + \frac{4\pi q^2}{k'^2} \sum_{\vec{k}} \left\{ \frac{f\left(\vec{k}\right) - f\left(\vec{k} + \vec{k'}\right)}{E\left(\vec{k} + \vec{k'}\right) - E\left(\vec{k}\right) - \hbar\omega + i\hbar\alpha} \right\} \tag{14.15}$$

This important formula is known as Lindhard's expression. The applied potential is $V(\vec{r},t)$, but the potential seen by the carriers is modified or screened by the medium to give $U(\vec{r},t)$ where:

$$V(\vec{r},t) = \iint d\vec{k'}\, d\omega e^{ik'.\vec{r}} e^{i\omega t} V(\vec{k'},\omega) \tag{14.16}$$

$$U(\vec{r},t) = \iint \frac{V(\vec{k'},\omega)}{\varepsilon(\vec{k'},\omega)} d\vec{k'}\, d\omega e^{i\vec{k'}.\vec{r}} e^{i\omega t} \tag{14.17}$$

14.2 Static Response

In order to appreciate the significance of this formula, consider the situation where the applied field is time independent, so we need to study $\varepsilon(k',0)$, i.e., at zero frequency. To do that we look at the limit $\vec{k'} \to 0$ in Eq. (14.15) where the denominator is largest and write:

$$\begin{aligned} E(\vec{k}+\vec{k'}) - E(\vec{k}) &= \vec{k'}.\nabla_{\vec{k}} E(\vec{k}) \\ f(\vec{k}) - f(\vec{k}+\vec{k'}) &= -\vec{k'}.\frac{\partial f_{\vec{k}}}{\partial E_{\vec{k}}} \nabla_{\vec{k}} E(\vec{k}) \end{aligned} \tag{14.18}$$

And in Eq. (14.15), we have:

$$\varepsilon(\vec{k'},0) \to 1 + \frac{q^2}{\varepsilon_0 k'^2} \int \frac{\vec{k'} \cdot \nabla_{\vec{k}} E(\vec{k})}{\vec{k'} \cdot \nabla_{\vec{k}} E(\vec{k})} \left(-\frac{\partial f}{\partial E}\right) d\vec{k} \tag{14.19}$$

$$\varepsilon(\vec{k'},0) = 1 + \frac{q^2}{\varepsilon_0 k'^2} \int \left(-\frac{\partial f}{\partial E}\right) g_V(E) dE \tag{14.20}$$

$$\varepsilon(\vec{k'},0) = 1 + \frac{\lambda_s^2}{k'^2} \tag{14.21}$$

where g_V is the density of states per unit volume. If we remember that $\left(-\frac{\partial f}{\partial E}\right)$ is at low temperatures almost a delta function at the Fermi energy, then this gives us:

$$\lambda_s = \frac{q^2 g_V(E_F)}{\varepsilon_0} \tag{14.22}$$

But the general result at any temperature follows from Eq. (14.20) and Eq. (14.21).

Now assume that the external potential is, for example, caused by an impurity with a Coulomb potential:

$$V(r) = \frac{q^2}{4\pi\varepsilon_0 r} \tag{14.23}$$

In Fourier space the bare Coulomb potential gives:

$$V(k') = \frac{q^2}{4\pi\varepsilon_0 k'^2} \tag{14.24}$$

so that the net potential seen by the other carriers in Fourier space is:

$$U = \left\{ \frac{q^2}{1 + \frac{\lambda_s^2}{k'^2}} \right\} \frac{1}{4\pi\varepsilon_0 k'^2} = \frac{q^2}{4\pi\varepsilon_0 \left(\lambda_s^2 + k'^2\right)} \tag{14.25}$$

And in real space this transforms back to:

$$U(r) = \frac{q^2}{4\pi\varepsilon_0 r} \exp[-\lambda_s r] \tag{14.26}$$

Now we understand that the quantity λ_s is an inverse screening length and depends on the magnitude of the density of states at the Fermi level. The density of states at the Fermi level is only finite when we have free carriers, i.e., when we have a finite conductivity at $T = 0$. In a metal, the screening length can be as short as $1/\lambda_s \sim 0.1$ nm. In a doped semiconductor, the screening length can be 100 times longer than that. Note that Eq. (14.21) is an approximation, and Eq. (14.26) is only valid at longer distances than the screening length. The exact evaluation and spatial dependence of the potential are quite a bit more complicated than that. For our purposes, however, the simple exponential result which is valid at long distances $r \gg 1/\lambda_s$ is good enough.

14.3 Screening in a Semiconductor

Now let us consider how charges are screened in a medium such as a semiconductor, where there are no free charges at low temperatures. For this purpose it is convenient to now explicitly index the bands so we have:

$$\varepsilon\left(\vec{k'},\omega\right) = 1 + \frac{4\pi q^2}{k'^2} \sum_{\vec{k};k',m} \left\{ \frac{\left| \left\langle \vec{k}, g \left| e^{i\vec{k'}\cdot\vec{r}} \right| \vec{k'}, m \right\rangle \right|^2 \left[f_g\left(\vec{k}\right) - f_m\left(\vec{k}+\vec{k'}\right) \right]}{\varepsilon_g\left(\vec{k}\right) - \varepsilon_m\left(\vec{k}+\vec{k'}\right) + \hbar\omega - i\hbar\alpha} \right\} \tag{14.27}$$

where g is the valence bands and m the conduction bands. The first allowed process in the sum is from the valence band to the conduction band, so we have for small $\vec{k'}$:

$$\varepsilon_m\left(\vec{k}+\vec{k'}\right) - \varepsilon_g\left(\vec{k}\right) \sim E_{\text{gap}} \tag{14.28}$$

In order to proceed further, we use the sum rule:

$$\sum_n (E_n - E_s) \left| \left\langle n \left| e^{i\vec{k}\cdot\vec{r}} \right| s \right\rangle \right|^2 = \frac{\hbar^2 k^2}{2m} \tag{14.29}$$

which follows from the commutator $[x,p] = i\hbar$. Then we use the approximation Eq. (14.28), assuming this is true for all k' to find:

$$\sum_m \left| \left\langle \vec{k}, g \left| e^{i\vec{k'}\cdot\vec{r}} \right| \vec{k}+\vec{k'}, m \right\rangle \right|^2 \sim \frac{\hbar^2 k'^2}{2mE_{\text{gap}}} \tag{14.30}$$

which then gives when substituted back into Eq. (14.27):

$$\varepsilon(k',0) = 1 + \frac{q^2}{\varepsilon_0 k'^2} \frac{n_v}{E_{\text{gap}}} \frac{\hbar^2 k'^2}{2m} \frac{2}{E_{\text{gap}}} \tag{14.31}$$

$$\varepsilon(k',0) = 1 + \frac{q^2}{\varepsilon_0} \frac{n_v}{E_{\text{gap}}} \frac{\hbar^2}{m} \frac{1}{E_{\text{gap}}} \tag{14.32}$$

$$\varepsilon(k',0) = 1 + \left(\frac{\hbar\omega_p}{E_{\text{gap}}} \right)^2 \tag{14.33}$$

where n_v is the density of electrons in the valence band and where we have defined the plasma frequency in the valence band as:

$$\omega_p = \left\{ \frac{n_v q^2}{m\varepsilon_0} \right\}^{1/2} \tag{14.34}$$

Going back to Eq.(13.14) again, and looking at the effective Coulomb potential, transforming back to real space, we now have:

$$V(r) = \frac{q^2}{4\pi\varepsilon_0\left[1 + \left(\frac{\hbar\omega_p}{E_G}\right)^2\right]r} \tag{14.35}$$

which, in contrast to Eq. (14.21), gives us a constant permittivity and keeps the long range nature of the Coulomb potential. The effective permittivity produced by the polarization of the valence band in the absence of free carriers is therefore:

$$\varepsilon_s = 1 + \left(\frac{\hbar\omega_p}{E_g}\right)^2 \tag{14.36}$$

The presence of free charges has a drastic effect on the screening as can be seen by comparing Eqs. (14.36) and (14.26). These results are some of the most important in solid-state physics.

According to Eq. (14.36), we have for direct bandgap zinc-blende materials, a valence band electron density which is the atom density, because we have four valence electrons per atom and thus a scaling of the permittivity with the energy gap. The smaller the gap, the larger the permittivity. Table 14.1 shows the experimental

Table 14.1 Table of important semiconductor parameters (see also Appendix A.4)

Semiconductor		Bandgap energy (eV)		Band	ε
		300 K	0 K		
Element	C	5.47	5.48	Indirect	5.7
	Si	1.12	1.17	Indirect	11.9
	Ge	0.66	0.74	Indirect	16.0
	Sn		0.082	Direct	
IV–IV	α-SiC	2.996	3.03	Indirect	10.0
III–V	BN	~7.5		Indirect	7.1
	GaN	3.36	3.50	Direct	12.2
	GaP	2.26	2.34	Indirect	11.1
	BP	2.0			
	AlSb	1.58	1.68	Indirect	14.4
	GaAs	1.42	1.52	Direct	13.1
	InP	1.35	1.42	Direct	12.4
	GaSb	0.72	0.81	Direct	15.7
	InAs	0.36	0.42	Direct	14.6
	InSb	0.17	0.23	Direct	17.7
II–VI	ZnS	3.68	3.84	Direct	5.2
	ZnO	3.35	3.42	Direct	9.0
	CdS	2.42	2.56	Direct	5.4
	CdSe	1.70	1.85	Direct	10.0
	CdTe	1.56		Direct	10.2
IV–VI	PbS	0.41	0.286	Indirect	17.0
	PbTe	0.31	0.19	Indirect	30.0

values for different semiconductors. Although there is a clear relationship, it is not as pronounced as that given by Eq. (14.36). There are two main reasons: the first one is that we have used quite a strong approximation in deriving Eq. (14.36). In particular the nature of the variations in the Bloch functions and effective masses has not been properly included in the evaluation of the permittivity and plasma frequencies. All that is left to distinguish the materials is the energy gap and the plasma frequency which itself also depends on the effective mass. The variations in the plasma frequency should of course also be included. Within the Kane model of Chap. 5, small m^* implies small bandgap and from Eq. (14.34) large plasma frequency, so the scaling with the energy gap is apparently even stronger. But the fact that this strong dependence is not observed has to do with the strong approximation we used to derive Eq. (14.33) which neglects the effect of the effective mass matrix elements. The second reason is that we have neglected exchange and correlation effects in discussing the electronic structure and assumed that one-body band theory is enough. Deriving the energy gap of semiconductors without these corrections turns out to be impossible. So it is not surprising that this simple scheme does not fully reproduce the experimental trend.

14.4 Screening in a 2-Dimensional System

The Lindhard function Eq. (14.15) is sensitive to the dimensionality of the system. For example, in a 2D free electron gas, Eq. (14.15) can be shown to become:

$$\varepsilon\left(\vec{k'},0\right) = 1 + \frac{q_{2d}}{k'}\left\{1 - \left[1 - \left(\frac{2k_F}{k'}\right)^2\right]^{1/2}\right\} \rightarrow k' > 2k_F \tag{14.37}$$

where:

$$q_{2d} = \frac{mq^2}{2\pi\varepsilon_0\varepsilon_b\hbar^2} = \frac{2}{a_{B,e}} \tag{14.38}$$

$$a_{B,e} = \frac{\pi\hbar^2\varepsilon_b\varepsilon_0}{mq^2} \tag{14.39}$$

and where we have assumed that the bound electrons give a constant permittivity ε_b or ε_s. This expression is actually very close to the classical 2D Thomas-Fermi function:

$$\varepsilon_{TF}\left(\vec{k'},0\right) = 1 + \frac{q_{2d}}{k'} \tag{14.40}$$

which gives us the screened Coulomb potential:

$$V_{2d}\left(\vec{r}\right)_{\mathrm{TF}} = \frac{q^2}{4\pi\varepsilon_0\varepsilon_b}\frac{q_{2d}(1+q_{2d}d)}{\left(rq_{2d}\right)^3} \tag{14.41}$$

and thus a cubic power law distance dependence. The quantum mechanical form Eq. (14.37) is more difficult to evaluate in real space, but at long distances, it has the interesting oscillatory structure:

$$V_{2d}(r)_{\mathrm{QM}} = -\frac{q^2}{4\pi\varepsilon_0\varepsilon_b}\frac{4\left(k_f\right)^2}{\left(2k_{\mathrm{F}}+q_{2d}\right)^2}\frac{\sin\left(2k_{\mathrm{F}}r\right)}{2k_{\mathrm{F}}r^2} \tag{14.42}$$

We note that the quantum mechanical result depends explicitly on the magnitude of the Fermi wave vector k_{F}, but the classical Thomas-Fermi result does not. The reader should also make a note of the very different screening properties of a 3D and 2D electron gases. This is very important in nanotechnology. The lower the dimensionality, the more ineffective the screening becomes. In one dimension, the Lindhard approximation is not accurate, so we have not discussed it here. One consequence is that in nanostructures, the effect of electron-electron interactions on the electronic transport and optical properties is much more significant than in the bulk. This has implications for engineering because it implies that physicists and engineers can use electron-electron interactions as an engineering design tool to find novel device functionalities.

14.5 Plasmon Modes

Consider again the 3D permittivity Eq. (14.15), and this time the very high-frequency limit:

$$\hbar\omega > \varepsilon\left(\bar{k}+\vec{k'}\right) - \varepsilon\left(\vec{k}\right) \tag{14.43}$$

with:

$$\varepsilon\left(\vec{k'},\omega\right) = 1 + \frac{q^2}{\varepsilon_0 k'^2}\sum_{\vec{k}}\left\{\frac{2f\left(\vec{k}\right)\left(E_{\bar{k}} - E_{\vec{k}+\vec{k'}}\right)}{-\left(E_{\vec{k}+\vec{k'}} - E_{\vec{k}}\right)^2 + (\hbar\omega)^2}\right\} \tag{14.44}$$

where Eq. (14.44) becomes after expanding for small $\vec{k'}$:

$$\varepsilon\left(\vec{k'},\omega\right) = 1 + \frac{q^2}{\varepsilon_0 k'^2}\sum_{\vec{k}}\frac{f_0\left(\vec{k}\right)}{(\hbar\omega)^2}\left(-k'^2\frac{\partial^2\varepsilon\left(\vec{k}\right)}{\partial k^2}\right) \tag{14.45}$$

giving:

$$\varepsilon(\omega,0)=1-\frac{\omega_p^2}{\omega^2}\rightarrow\omega_p^2=\frac{q^2n}{m\varepsilon_0}(\text{MKS}) \tag{14.46}$$

The reader should note that we have encountered this result before when we were analyzing the permittivity of the electron gas in Drude theory; see Eq. (14.37).

If we do the same analysis in 2D, we obtain a different plasma dispersion. We quote here the result which is (q = electron charge):

$$\omega_{p,2D}\left(k_p\right)=\sqrt{\frac{q^2n_{2D}k_p}{2\varepsilon_0\varepsilon_b m}} \tag{14.47}$$

We note that the zero wavevector value of the 2D plasmon dispersion is zero in contrast to the 3D result. Plasmon modes are very much geometry and system size dependent. The general plasmon dispersion relation can be obtained from the requirement that there be longitudinal mode solutions in the Maxwell equations and thus that the wavevector and frequency-dependent permittivity be:

$$\varepsilon\left(\vec{k},\omega\right)=0 \tag{14.48}$$

14.6 Surface Plasmons

Plasmon modes or, in other words, the collective oscillations of electron clouds in bulk 2D systems and nanoparticles are the so-called surface plasmons. This is a field of great current interest, and the reader is referred to the specialized literature on this topic available from Internet searches. The point is that, normally, in solid-state physics, we consider the properties and response behavior of single charges. But in general and especially when we apply a time-dependent electric field, we also have to take account of the fact that all the free electrons experience the same stimulation and therefore produce internal and external responses which are the result of the motion of many charges. This collective response can be very much larger than the response of a single particle and enhance the total electric field seen by individual charges. Thus if we place a charge near a small spherical metal particle and then apply an oscillating field to the system, the charge sees not only the applied field but also the field produced by the collective motion or response of the free electrons in the metallic nanoparticle (see Pinchuk et al. 2004, in the references). This additional field would not be very significant until the frequency of the applied stimulation reaches the plasma frequency of the nanoparticle. When this happens, i.e., at resonance, the collective response becomes very large and can be many orders of magnitude bigger than the original stimulating field. Clearly these types of processes imply many novel applications. One can, in this way, enhance local fields by using surface plasmon amplifiers, by many orders of magnitude, and thus is a very topical

field in modern solid-state engineering. The reader is referred to the specialized literature on the subject.

14.7 Summary

In this chapter we have in some sense completed the work we started in Chap. 10 and investigated how bound electrons and free electrons change the electric fields that charges produce in solids. This gave rise to the concept of screening which we have in part already encountered in elementary electromagnetic theory. We derived the important Lindhard function. We noted how drastically a Coulomb potential is modified at long range by the presence of free electrons. This is ultimately one of the reasons which one-body approximations work so well in solids. Then we discovered that the screening is strongly dimensionality dependent becoming less and less effective the lower the dimensionality of the system. This has a strong impact on "nanotechnology" and makes electron-electron interaction a serious design tool. For example, a single trapped charge can block an entire current path in a thin enough wire. Removing the charge will open the channel again. We also briefly touched on the exciting new field of "surface plasmon" research and development and urged the interested reader to consult the specialized literature.

Problems

1. Explain what is meant by screening of electrical potentials. Explain the difference between the screening properties of metals and insulators. If in a solid the density of states at the Fermi level $g_v(E_F)$ is $10^{26}/m^3eV$, what is the screening length? If we lower the temperature and the solid turns into an insulator or wide bandgap semiconductor, what happens to the screening length?
2. What is a plasmon? What is the plasma frequency of a 3D metal for which the electron density is $n_c = 10^{27}/m^3$. If you were asked to choose materials or design a system for which the plasmon frequency is in the regime of $\hbar\omega_p \sim 0.5$ eV, what would you choose? What is the free electron density needed to obtain such a plasmon frequency?

References

Pinchuk A, von Plessen G, Kreibig U (2004) Influence of interband electronic transitions on the optical absorption in metallic nanoparticles. J Phys D Appl Phys 37:3133

Ziman JM (1964) The principles of the theory of solids. Cambridge University press, Cambridge

Further Reading

Ando T, Folwler AB, Stern F (1982) Electronic properties of two dimensional systems. Rev Mod Phys 54:437

Ahmed H (1986) An integration microfabrication system for low dimensional structures and devices. In: Kelly MJ, Weisbuch C (eds) The physics and fabrication of microstructures and microdevices. Springer, Berlin, pp 435–442

Asada M, Miyamoto Y, Suematsu Y (1986) Gain and the threshold of 3-dimensional quantum-box lasers. IEEE J Quantum Electron 22:1915–1921

Ashcroft NW, Mermin ND (1976) Solid state physics. Holt, Rinehart, Winston, New York

Bassani F, Pastori Parravicini G (1975) Electronic states and optical transitions in solids, chap. 6. Pergamon, New York

Bastard G (1988) Wave mechanics applied to semiconductor heterostructures. Halsted Press, New York

Beaumont SP (1992) Quantum wires and dots-defect related effects. Phys Scr T45:196–199

Bockrath M, Cobden DH, Lu J, Rinzler AG, Smalley RE, Balents L, McEuen PL (1999) Luttinger liquid behaviour in carbon nanotubes. Lett Nat 397:598–601

Chuang SL (1995) Physics of optoelectronic devices. John Wiley & Sons, New York

Davies JH (1998) The physics of low dimensional semiconductors: an introduction. Cambridge University Press, New York

Dingle R (1975) Confined carrier quantum states in ultrathin semiconductor heterostructures. In: Queisser HJ (ed) Feskorperproblem XV. Springer, pp 21–48

Dingle R, Wiegmann W, Henry CH (1974) Quantum states of confined carriers in very thin $Al_xGa_{1-x}As$-GaAs- $Al_xGa_{1-x}As$ heterostructures. Phys Rev Lett 33:827–830

Einspruch NG, Frensley WR (1994) Heterostructures and quantum devices. Academic Press Limited, London

Hasko DG, Potts A, Cleaver JR, Smith C, Ahmed H (1988) Fabrication of sub-micrometer free standing single crystal GaAs and Si structures for quantum transport studies. J Vac Sci Technol B 6:1849–1851

Kelly MJ (1995) Low-dimensional semiconductors: materials, physics, technology, devices. Oxford University Press, New York

Mohseni H (2001) *Type-II InAs/GaSb superlattices for infrared detectors*, Thesis, Ph.D. dissertation, Northwestern University

Razeghi M, Duchemin JP, Portal JC, Dmowski L, Remeni G, Nicolas RJ, Briggs A (1986) First observation of the quantum hall effect in a $Ga_{0.47}In_{0.53}As$-InP heterostructure with three electric subbands. Appl Phys Lett 48:712–715

Rosencher E, Vinter B (2002) Optoelectronics. Cambridge University Press, Cambridge

Scherer A, Jewell J, Lee YH, Harbison J, Florez LT (1989) Fabrication of microlasers and microresonator optical switches. Appl Phys Lett 55:2724–2726

Shockley W (1951) U.S. Patent 2,569,347

Sze S (1981) Physics of semiconductor devices, 2nd edn. John Wiley & Sons, New York

Tewordt M, Law V, Kelly M, Newbury R, Pepper M, Peacock C (1990) Direct experimental determination of the tunneling time and transmission probability of electrons through a resonant tunneling system. J Phys Condens Matter 2:896–899

Vasko FT, Kuznetsov AV (1999) Electronic states and optical transitions in semiconductor heterostructures. Springer, New York

Weiner JS, Miller DAB, Chemla DJ, Damen TC, Burrus CA, Wood TH, Gossard AC, Wiegmann W (1985) Strong polarization sensitive electroabsorption in GaAs/AlGaAs quantum well waveguides. Appl Phys Lett 47:1148–1151

Weisbuch C, Vinter B (1991) Quantum semiconductor structures. Academic Press, New York

Semiconductor Heterostructures and Low-Dimensional Quantum Structures

15

15.1 Introduction

In Chap. 4, we have introduced the basic concepts and formalism of quantum mechanics. In Chap. 5, we have determined the energy spectrum, or energy-momentum or E–k relations, for electrons in a crystal which governs their interaction with external forces and fields. Moreover, we saw that the quantum behavior of particles is best observed in small, typically nanometer scale (one billionth of a meter or 10^{-9} m) dimension structures, as illustrated in the example of a particle in a 1D box.

In nanometer-scale structures in a crystal, the motion of an electron can be confined in one or more directions in space. When only one dimension is restricted while the other two remain free, we talk about a quantum well; when two dimensions are restricted, we talk about a quantum wire; and when the motion in all three dimensions is confined, we talk about a quantum dot. In solid-state engineering, these are commonly called *low-dimensional quantum structures*.

In the past few decades, progress in semiconductor crystal growth technology, such as liquid-phase epitaxy (LPE), molecular beam epitaxy (MBE), and metal-organic chemical vapor deposition (MOCVD), has made it possible to control with atomic-scale precision of the dimensions of semiconductor structures and thus to realize such low-dimensional quantum structures through the formation of heterojunctions or heterostructures. A semiconductor heterojunction is formed when two different semiconducting materials are brought into direct contact with each other, while heterostructures can be defined as materials that incorporate one or more heterojunctions and can describe more complicated device architectures such as multiple quantum wells, superlattices, and other low-dimensional quantum structures.

First proposed by Shockley in 1951 in a heterojunction bipolar transistor (HBT) (Shockley 1951), heterojunctions have been used heavily in a variety of applications. Many conventional devices take advantage of the special properties of

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_15

heterostructures including semiconductor lasers, light-emitting diodes, photodetectors, etc.

There exist several inherent design advantages to using heterojunctions as opposed to standard homojunctions in semiconductor devices. Due to pairing small- and wide-bandgap materials or by tailoring their lineup energy position, charge carriers can be confined or redistributed. This offers the chance to control, to considerable extent, the physical location of free electrons and holes within the device as well as the wavefunction overlap between the carrier types. Furthermore, by choosing the semiconducting materials and the doping level, important properties of the heterostructure device can be designed. This includes the bandgap, the effective mass, and the carrier transport. Finally, depending on the lattice mismatch between the heterojunction materials, built-in strain fields can be engineered and used to obtain enhanced electrical or optical properties.

This chapter will first review the concepts associated with semiconductor heterostructures, including energy band offsets, types of alignment, and a few models for heterojunction energy band alignment. Then, the properties of low-dimensional quantum structures will be discussed in detail.

15.2 Energy Band Offsets

When a heterojunction is formed, the conduction and valence band alignment is dependent upon the properties of the constituent materials such as their bandgap, the doping, and the electron affinity. Heterostructures can be classified depending on the band alignment formation between the two semiconductor materials. The possible band alignment combinations include “type I,” “type II staggered,” and “type II broken gap” and are described below.

15.2.1 Type I Alignment

When the valence and conduction band of one material “straddles” the bands of the narrow-gap material, the heterojunction band alignment is termed type I. The heavily investigated AlGaAs/GaAs heterojunction exhibits this band lineup with the aluminum-containing material having its conduction band above and valence band below the corresponding GaAs band energies. An example of type I band alignment is shown in Fig. 15.1a. The schematic figure shows materials in electrical isolation from one another. As we will see later in this chapter, direct interaction between semiconductor materials results in space-charge redistribution, which leads to band bending near the junction position.

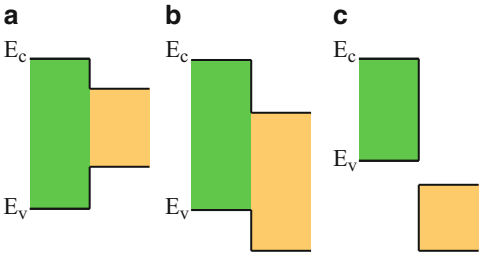


Fig. 15.1 Heterojunction band lineups for isolated but adjacent semiconductors: (**a**) type I, (**b**) type II staggered, and (**c**) type II broken gap alignments

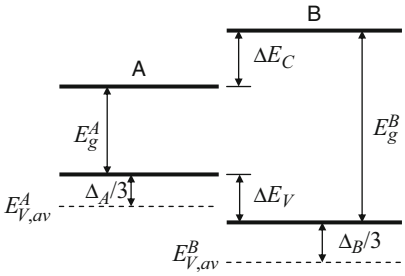


Fig. 15.2 Band alignment diagram for calculation of band offset

15.2.2 Type II Alignments

Semiconductor heterojunctions may also form where the conduction and valence bands in one material are both slightly below the corresponding band energies in the adjacent semiconductor. This band alignment is termed type II staggered and is shown in Fig. 15.1b. One example of a heterojunction material system that can be generally classified as type II staggered is InAs/AlSb.

The InAs/GaSb heterojunction is an example of a type II broken gap alignment. This occurs when the conduction of one material is at a lower energy than the valence band of the adjacent semiconductor. An example of broken-gap band alignment is shown in Fig. 15.1c.

15.3 Application of Model Solid Theory

In the previous sections, we have introduced different types of band lineups. In order to better understand the heterojunction properties, it is important to determine the actual band lineups between two different materials. We introduce the application of model solid theory for this type of calculation. For simplicity, we consider unstrained junctions only. This is true for the GaAs/$Al_xGa_{1-x}As$ ($0 < x < 0.4$) junction system.

We assume A and B in Fig. 15.2 represent two III–V semiconductors that have the same lattice constant. The valence band position can be calculated as:

$$E_V = E_{V,av} + \frac{\Delta}{3} \tag{15.1}$$

in which $E_{V,av}$ is the average valence band position which is obtained from theory and is referred to as the absolute energy level, E_V is the valence band position, and Δ is the spin-orbit splitting energy. The values for different semiconductors are usually tabulated in the literature.

The valence band offset between semiconductor A and B thus can be calculated as:

$$\Delta E_V = \left(E_{V,av}^A - E_{V,av}^B\right) + \frac{1}{3}(\Delta_A - \Delta_B) \tag{15.2}$$

The conduction band edge is obtained by adding the bandgap value to the valence band position:

$$E_C = E_V + E_g \tag{15.3}$$

Therefore the conduction band offset can be calculated as:

$$\Delta E_C = \left(E_{V,av}^A - E_{V,av}^B\right) + \frac{1}{3}(\Delta_A - \Delta_B) + \left(E_g^A - E_g^B\right) \tag{15.4}$$

All these quantities are summarized in Fig. 15.2.

Example
Q: Determine the band offset of a GaAs/$Al_{0.2}Ga_{0.8}As$ heterojunction. The material parameters for GaAs and AlAs are listed in Table 15.1.
A: For GaAs, we have:

$$E_V^{GaAs} = E_{V,av}^{GaAs} + \frac{\Delta_{GaAs}}{3} = -6.807\text{eV}$$

For $Al_{0.2}Ga_{0.8}As$, we use the arithmetic average of 20% AlAs and 80% of GaAs:

$$\begin{aligned} E_V^{Al_{0.2}Ga_{0.8}As} &= 0.2 \times \left(E_{V,av}^{AlAs} + \frac{\Delta_{AlAs}}{3}\right) + 0.8 \times \left(E_{V,av}^{GaAs} + \frac{\Delta_{GaAs}}{3}\right) \\ &= -6.925\text{eV} \end{aligned}$$

$$E_g^{Al_{0.2}Ga_{0.8}As} = 0.2 \times E_g^{AlAs} + 0.8 \times E_g^{GaAs} = 1.842\text{eV}$$

Table 15.1 Material parameters for GaAs and AlAs

	$E_{V,av}$ (eV)	Δ(eV)	E_g(eV)
GaAs	−6.92	0.34	1.52
AlAs	−7.49	0.28	3.13

Therefore, we obtain the band offset as follows:

$$\begin{cases} \Delta E_{\mathrm{V}} = E_{\mathrm{V}}^{\mathrm{GaAs}} - E_{\mathrm{V}}^{\mathrm{Al_{0.2}Ga_{0.8}As}} = (-6.807) - (-6.925) \\ \qquad = 0.118\mathrm{eV} \\ \Delta E_{\mathrm{C}} = \left(E_{\mathrm{V}}^{\mathrm{Al_{0.2}Ga_{0.8}As}} + E_{\mathrm{g}}^{\mathrm{Al_{0.2}Ga_{0.8}As}}\right) - \left(E_{\mathrm{V}}^{\mathrm{GaAs}} + E_{\mathrm{g}}^{\mathrm{GaAs}}\right) \\ \qquad = (-5.287) - (-5.555) \\ \qquad = 0.268\mathrm{eV} \end{cases}$$

15.4 Anderson Model for Heterojunctions

When we bring two different semiconductors in contact with each other, due to their difference of the Fermi level with respect to the vacuum level, there will be net charge transfer from one material to the other. At equilibrium, the Fermi level lines up on both sides of the junction. This will change the band diagram of the heterojunction from straight lines to partially rounded curves. In this section, we use the basic Anderson model to calculate the zero-bias band diagram for a *p-n* junction made from a type I heterojunction, with N_{A} representing the *p*-type doping level of the narrower-gap material and N_{D} the *n*-type doping level of the wider-gap material. The other cases of *p-n* heterojunctions can be derived in a same manner and will not be covered.

To simplify the calculations and emphasize the methodology that will be introduced, we assume that both N_{A} and N_{D} are much larger than the intrinsic carrier concentration and that all the dopants are ionized. Before contact, the Fermi level on each side is represented as $E_{\mathrm{F_A}}$ and $E_{\mathrm{F_B}}$. We use V_0 to represent the potential difference due to the energy difference between $E_{\mathrm{F_A}}$ and $E_{\mathrm{F_B}}$, as shown in Fig. 15.2. According to Fig. 15.2, we have:

$$V_0 = E_{\mathrm{g}}^{\mathrm{A}} + \Delta E_{\mathrm{C}} - \left(E_{\mathrm{F_A}} - E_{\mathrm{V}}^{\mathrm{A}}\right) - \left(E_{\mathrm{C}}^{\mathrm{B}} - E_{\mathrm{F_B}}\right) \tag{15.5}$$

For nondegenerate semiconductors, we have:

$$\begin{cases} E_{\mathrm{F_A}} - E_{\mathrm{V}}^{\mathrm{A}} = -k_{\mathrm{b}} T \ln\left(\dfrac{N_{\mathrm{A}}}{N_{\mathrm{v}}^{\mathrm{A}}}\right) \\ E_{\mathrm{V}}^{\mathrm{B}} - E_{\mathrm{F_B}} = -k_{\mathrm{b}} T \ln\left(\dfrac{N_{\mathrm{d}}}{N_{\mathrm{c}}^{\mathrm{B}}}\right) \end{cases} \tag{15.6}$$

where $N_{\mathrm{v}}^{\mathrm{A}}$ and $N_{\mathrm{c}}^{\mathrm{B}}$ are the valence band and conduction-band density of states for semiconductor A and B, respectively. Substituting Eq. (15.6) into Eq. (15.5), we obtain the expression for V_0:

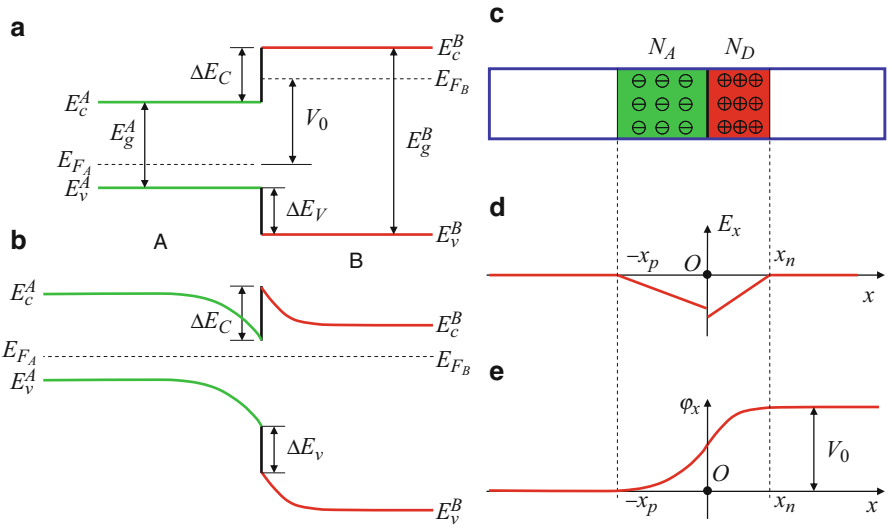


Fig. 15.3 Illustrations for (**a**) band diagram for the heterojunction before charge transfer, (**b**) band diagram after charge transfer, (**c**) depletion approximation, (**d**) electric field distribution, and (**e**) electrical potential distribution

$$V_0 = E_{\mathrm{g}}^{\mathrm{A}} + \Delta E_{\mathrm{C}} + k_{\mathrm{b}} T \ln\left(\frac{N_{\mathrm{A}} \cdot N_{\mathrm{D}}}{N_{\mathrm{v}}^{\mathrm{A}} \cdot N_{\mathrm{c}}^{\mathrm{B}}}\right) \tag{15.7}$$

After we bring semiconductor A and B together into contact, there will be a net electron transfer from B to A (see Fig. 15.3c) until the Fermi levels on both sides reach the same value, as shown in Fig. 15.3b.

The number of excess negative charges (ionized acceptors) on the *p*-side will be exactly the same as that of the excess positive charges (ionized donors) on the n-side. N_{a} and N_{d} equal the charge densities on the p and n-sides of the junction within the depletion region. Thus we have the charge conservation equation:

$$N_{\mathrm{A}} x_p = N_{\mathrm{D}} x_n \tag{15.8}$$

We assume that the charge density is uniformly distributed on either side of the junction over a certain distance. This is called the depletion approximation. Under this approximation, we can calculate the electric field distribution and thus the electrical potential profile.

Assume that ε_{A} and ε_{B} represent the relative permittivity for semiconductor A and B. Using Gauss' law, we can obtain the electric field within the depletion region as:

$$\begin{cases} E_x = -\dfrac{qN_{\mathrm{A}}(x + x_p)}{\varepsilon_{\mathrm{A}}\varepsilon_0}, & -x_p \le x < 0 \\ E_x = -\dfrac{qN_{\mathrm{D}}(x_n - x)}{\varepsilon_{\mathrm{B}}\varepsilon_0}, & 0 < x \le x_n \end{cases} \tag{15.9}$$

Outside the depletion region, the net charge density is zero, and there is no electric field. We take the zero potential to be at the neutral region in the semiconductor A. We integrate the electric field from the point of calculation toward the potential zero point to obtain the electrical potential profile:

$$\varphi_x = \int_x^{-x_p} E_x dx \tag{15.10}$$

Substituting Eq. (15.9) into Eq. (15.10), we have:

$$\begin{cases} \varphi_x = 0, & x < -x_p \\ \varphi_x = \dfrac{qN_{\mathrm{A}}\left(x + x_p\right)^2}{2\varepsilon_{\mathrm{A}}\varepsilon_0}, & -x_p \le x < 0 \\ \varphi_x = \dfrac{qN_{\mathrm{A}}x_p^2}{2\varepsilon_{\mathrm{A}}\varepsilon_0} + \dfrac{qN_{\mathrm{D}}(2x_n x - x^2)}{2\varepsilon_{\mathrm{B}}\varepsilon_0}, & 0 \le x \le x_n \\ \varphi_x = \dfrac{qN_{\mathrm{A}}x_p^2}{2\varepsilon_{\mathrm{A}}\varepsilon_0} + \dfrac{qN_{\mathrm{D}}x_n^2}{2\varepsilon_{\mathrm{B}}\varepsilon_0}, & x > x_n \end{cases} \tag{15.11}$$

We recall that the total potential drop is V_0 as calculated before, i.e.,

$$\frac{qN_{\mathrm{A}}x_p^2}{2\varepsilon_{\mathrm{A}}\varepsilon_0} + \frac{qN_{\mathrm{D}}x_n^2}{2\varepsilon_{\mathrm{B}}\varepsilon_0} = V_0 \tag{15.12}$$

Combining Eq. (15.8) and Eq. (15.12), we obtain the values of x_{n} and x_{p} in terms of V_0:

$$\begin{cases} x_n = \sqrt{\dfrac{N_{\mathrm{A}}}{N_{\mathrm{D}}} \dfrac{2\varepsilon_0 V_0}{q} \dfrac{\varepsilon_{\mathrm{A}}\varepsilon_{\mathrm{B}}}{N_{\mathrm{A}}\varepsilon_A + N_{\mathrm{D}}\varepsilon_{\mathrm{B}}}} \\ x_p = \sqrt{\dfrac{N_{\mathrm{D}}}{N_{\mathrm{A}}} \dfrac{2\varepsilon_0 V_0}{q} \dfrac{\varepsilon_{\mathrm{A}}\varepsilon_{\mathrm{B}}}{N_{\mathrm{A}}\varepsilon_{\mathrm{A}} + N_{\mathrm{D}}\varepsilon_{\mathrm{B}}}} \end{cases} \tag{15.13}$$

We define the junction depletion width as $x_{\mathrm{w}} = x_{\mathrm{n}} + x_{\mathrm{p}}$. Taking into account Eq. (15.13), we can obtain:

$$x_w = \sqrt{\frac{2\varepsilon_0 V_0}{qN_{\mathrm{D}}N_{\mathrm{A}}} \frac{\varepsilon_{\mathrm{A}}\varepsilon_{\mathrm{B}}}{N_{\mathrm{A}}\varepsilon_{\mathrm{A}} + N_{\mathrm{D}}\varepsilon_{\mathrm{B}}}} \cdot (N_{\mathrm{D}} + N_{\mathrm{A}}) \tag{15.14}$$

Substituting Eq. (15.13) into Eq. (15.12), we will obtain the values for the electrical potential φ_x. In order to update the electron energy band diagram, we need to take into account that the electron charge is negative and the electron energy profile will be inverted. Adding this energy profile to the flat band profile as shown in Fig. 15.3a, we will obtain a calculated electron energy profile for the heterojunction under equilibrium as illustrated in Fig. 15.3b.

15.5 Multiple Quantum Wells and Superlattices

By "sandwiching" a low-bandgap material between two layers of wider bandgap material, a device designer can fabricate a single quantum well, as discussed later in this chapter. A layer of GaAs between two $Al_xGa_{1-x}As$ barriers acts as a potential well for electrons and holes. By adjusting the well width and composition of the barriers, one can engineer specific properties into the quantum well structure such as the energy bandgap.

In a similar fashion, multiple quantum wells (MQWs) may be formed by epitaxy of successive, periodic heterojunctions. Typically within MQWs, the carriers within a quantum well do not interact with carriers in a neighboring well. In other words, the electron and hole wavefunctions between adjacent wells do not overlap. Depending on the band alignment type of the heterojunctions involved, electrons and holes can be confined in similar or different spatial locations in the multiple quantum well structure. Multiple quantum wells are used in devices like quantum well intersubband photodetectors (QWIP) for enhanced absorption over a thicker active region.

Superlattices are structures that also have periodic heterojunctions similar to multiple quantum wells. However, the confined charge carriers within the individual quantum wells actively interact with carriers in other wells. This can be achieved by decreasing the quantum well barrier thickness in a multiple quantum well structure. The electron is now delocalized and can move from well to well just as in a Kronig-Penney lattice. Over an extended length span (many superlattice periods), electrons in superlattices can therefore exhibit miniband behavior, similar to bulk crystals. By controlling the layer structure, the superlattice band structure can be engineered. One can enhance desired effects such as optical emission/absorption or reduce unwanted effects such as Auger recombination. In addition, properties such as tunneling transport can be modified. An example of an epitaxially grown InAs/GaSb superlattice is shown in Fig. 15.4.

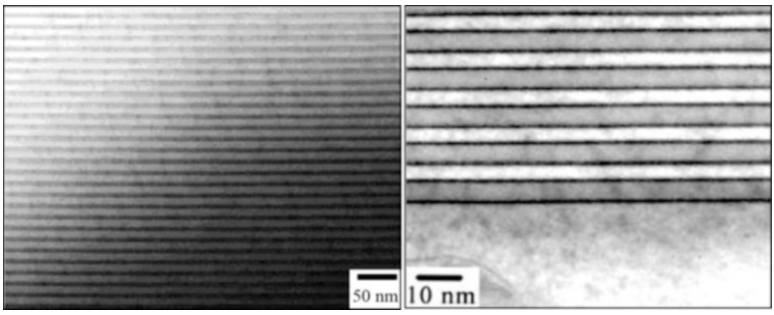


Fig. 15.4 Transmission electron microscope images of type II InAs/GaSb superlattice. The dark regions correspond to the InSb interface between InAs and GaSb layers (by courtesy of G Brown)

15.6 Two-Dimensional Structures: Quantum Wells

15.6.1 Energy Spectrum

As briefly mentioned previously, a quantum well is formed when the motion of electrons is confined in one direction (e.g., x), while it remains free to move in the other two directions (y, z). This situation is most easily achieved by sandwiching a thin and flat film semiconductor crystal between two crystals of another other semiconductor material in such a way that a potential step is produced, as shown in Fig. 15.5. The electrons are confined in the region $0 < x < a$. In the following discussion, we chose U_0, the potential step, to be finite.

This energy profile is in fact a potential that an electron experiences when moving through the structure. This is in addition to the crystal periodic potential of Chap. 3, which will not be brought into the discussion as it is already taken into account by considering an effective mass for the electron.

The potential in the x-direction is analogous to the case of a particle in a finite potential well as discussed in Sect. 4.4.4. The height of the potential barrier is now the difference between the conduction band energies in the different semiconductors, which is called the conduction band offset and denoted ΔE_c. The contribution to potential in the y- and z-directions is constant and is chosen to be zero, similar to the case of a free particle, as discussed in Sect. 4.4.1. The total potential can therefore be expressed as:

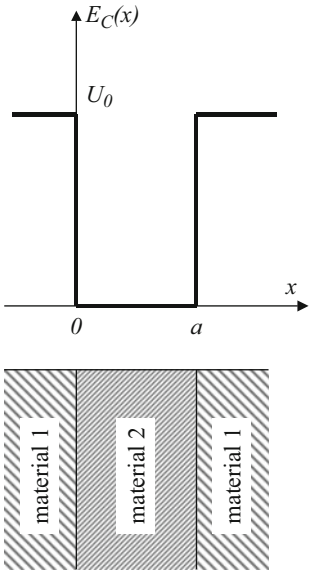


Fig. 15.5 Potential energy profile of a quantum well. This profile can be obtained by sandwiching a thin and flat semiconductor film of material 2 between two semiconductor crystals of another material 1

$$U(x,y,z) = \begin{cases} 0 & \text{for } 0 < x < a \\ U_0 > 0 & \text{for } x < 0 \text{ and } x > a \end{cases}, \tag{15.15}$$

and the time-independent Schrödinger equation becomes:

$$-\frac{\hbar^2}{2m^*}\nabla^2\Psi(x,y,z) - [E - U(x,y,z)]\Psi(x,y,z) = 0 \tag{15.16}$$

where m^* is the electron effective mass. The shape of the potential in Eq. (15.15) implies that the motion in the x-direction and that in the (y,z)-plane are independent. It is a common practice to use the subscripts "$\perp$" and "//" to denote the motion and energies for the x-direction and (y,z)-plane, respectively. For example, $\vec{r}$ is used to denote the position vector in the x-direction and $\vec{r}_{//}$ the position vector in the (y,z)-plane. The total three-dimensional wavefunction can therefore be represented by the product of two functions, one dependent on x alone and the other on (y,z) only, $\Psi_{\text{total}}(x,y,z) = \Psi_{//}(\vec{r}_{//})\Psi_{\perp}(\vec{r}_{\perp})$, and the total energy spectrum consists of the sum of two independent contributions: $E(\vec{r}) = E_{//}(\vec{k}_{//}) + E_{\perp}(\vec{k}_{\perp})$. Now let us consider the wavefunctions and energy spectrum in more detail.

In-Plane Motion

In the (y,z)-plane, the motion of the electron is similar to that of a free particle discussed in Sect. 4.4.1. The wavefunction $\Psi_{//}(\vec{r}_{//})$ can therefore be considered to be a plane wave similar to Eq. (4.32) and can be expressed as:

$$\Psi_{//}(\vec{r}_{//}) = A\exp(i\vec{k}_{//}\cdot\vec{r}_{//}) \tag{15.17}$$

where A is a normalization constant. The energy spectrum in the (y,z) plane is given by:

$$E_{//}(\vec{k}_{//}) = \frac{\hbar^2\vec{k}_{//}^{\,2}}{2m^*} = \frac{\hbar^2\left(k_y^2 + k_z^2\right)}{2m^*} \tag{15.18}$$

Note that these expressions are correct only for small values of the momentum such that $\left|\vec{k}_{//}\right| \ll \left|\vec{K}\right|$, where $\vec{K}$ is a reciprocal lattice vector. This restriction arises from the fact that we are not considering a completely free particle but rather an electron in a crystal. For a more precise discussion on what happens near a reciprocal lattice vector, the reader may be referred to the Kronig-Penney model in Chap. 5.

Motion Perpendicular to Well Plane

In the x-direction, the discussion is the same as that of a particle in a finite potential well conducted in Sect. 4.4.4. Although no analytical solution was derived, the main results can be summarized as follows.

The set of equations from Eq. (4.57) to Eq. (4.59) yields the quantized allowed energy levels $E_{\perp n}$, momenta $k_{\perp n}$, and decay coefficients α_n for an electron in this potential well, indexed by an integer $n = 0,1,\ldots$, and these quantities must satisfy Eq. (4.50):

$$\begin{cases} E_{\perp n} = \dfrac{\hbar^2 (k_{\perp n})^2}{2m^*} \\ \alpha_n = \sqrt{\dfrac{2m^*(U_0 - E_{\perp n})}{\hbar^2}} \end{cases} \tag{15.19}$$

Note that we are now using the effective mass of the electron, m^*. The spacing between consecutive energy levels is on the order of $\hbar^2\pi^2/m^*a^2$ from Eq. (4.43). For $E_n < U_0$, the wavefunction $\Psi_\perp\left(\vec{r}_\perp\right) = \Psi_\perp(x)$ consists of an oscillatory function inside the well ($0 < x < a$) and a decaying exponential outside the well. If needed, this wavefunction can be calculated using Eq. (4.51), Eq. (4.53), and the values of $E_{\perp n}$, $k_{\perp n}$, and α_n as illustrated in Fig. 15.6.

For an electron in a perfect crystal, the quantization of the energy levels and momenta is significant only when the dimensions of the confining structure (e.g., a) become on the order of or less than the electron de Broglie wavelength (Eq. (4.1)).

In a real crystal, however, there are defects which introduce perturbations of the potential periodicity. This results in the broadening of the initially discrete energy levels, and the magnitude of this broadening can be estimated to be $\hbar/\tau$ where τ is a characteristic time between electron collisions, or electron lifetime, which can be

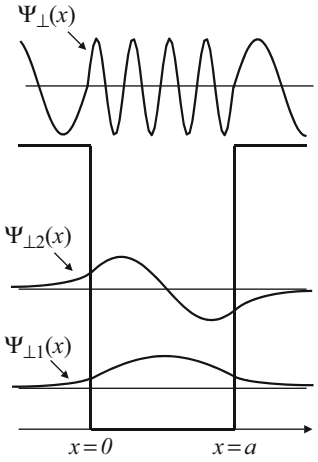


Fig. 15.6 Shapes of the wavefunctions $\Psi_\perp(x)$ for the allowed energy levels of a quantum well. In this example, there are only two allowed confined states. The wavefunctions of these have an oscillatory behavior inside the well ($0 < x < a$) but vanish rapidly when outside the quantum well. A third allowed state is shown which has an energy above the barrier of the well and therefore corresponds to a non-confined state. Its wavefunction has an oscillatory behavior in the entire space

understood as the average duration between two consecutive encounters with defects. A detailed discussion on electron collisions is beyond the scope of this textbook, and the reader is referred to the Further Reading section.

In such a situation, the quantization of the energy levels can be resolved only if the energy spacing between consecutive levels ($\hbar^2\pi^2/m^*a^2$) is larger than the broadening ($\hbar/\tau$). In other words, the inequality $\hbar^2\pi^2/m^*a^2 \gg \hbar/\tau$ ensures that the quantized behavior can be observed.

15.6.2 Density of States

The total energy spectrum for an electron in a quantum well is given by considering Eq. (15.18) and Eq. (15.19):

$$E\left(\vec{k}_{//}, n\right) = E_{//}\left(\vec{k}_{//}\right) + E_{\perp n} = \frac{\hbar^2\,\vec{k}_{//}^{\,2}}{2m^*} + \frac{\hbar^2(k_{\perp n})^2}{2m^*} \tag{15.20}$$

where the values of $\vec{k}_{//}$ are continuous while $k_{\perp n}$ is quantized and indexed by an integer n. Similar to Eq. (5.41) in Sect. 5.3.2, the density of states for quasi-two-dimensional electrons in quantum well is the number of allowed electron energy states (taking into account spin degeneracy) per unit energy interval around an energy E and is given by:

$$g_{2\mathrm{D}}(E) = 2\sum_{n,\,\vec{k}_{//}} \delta\left[E_{//}\left(\vec{k}_{//}\right) + E_{\perp n} - E\right] \tag{15.21}$$

where the factor 2 arises from the spin degeneracy. In this case, because one dimension is quantized while the other two remain continuous, the summation in Eq. (5.44) is performed on two coordinates only:

$$\sum_{\vec{k}_{//}} Y\left(\vec{k}_{//}\right) = \frac{S}{(2\pi)^2}\int\int Y\left(\vec{k}_{//}\right)d\vec{k}_{//} = \frac{S}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Y\left(k_y, k_z\right)dk_y dk_z \tag{15.22}$$

where S is the cross-sectional area of the crystal, in the (y,z)-plane, and $Y\left(\vec{k}_{//}\right)$ is an arbitrary function. Equation (15.21) then becomes:

$$g_{2\mathrm{D}}(E) = \frac{2S}{(2\pi)^2}\sum_{n}\iint_{\vec{k}_{//}} \delta\left[E_{//}\left(\vec{k}_{//}\right) + E_{\perp n} - E\right]d\vec{k}_{//} \tag{15.23}$$

Now, we must determine the relation between $d\left[E_{//}\left(\vec{k}_{//}\right)\right]$ as a function of $d\,\vec{k}_{//}$ in order to perform the integration in Eq. (15.23). For this, we follow the same analysis conducted in Eq. (5.35) to Eq. (5.39). Equation (15.23) yields:

$$d\left[E_{//}\left(\vec{k}_{//}\right)\right] = \frac{\hbar^2}{2m^*}.(2k_{//})dk_{//} \tag{15.24}$$

where $k_{//}$ is the norm or length of the vector $\vec{k}_{//}$. On the other hand, in two dimensions, Eq. (5.37) becomes:

$$d\vec{k}_{//} = d\left(\pi k_{//}^2\right) = 2\pi k_{//}dk_{//} \tag{15.25}$$

Thus:

$$d\left[E_{//}\left(\vec{k}_{//}\right)\right] = \frac{\hbar^2}{2m^*}\frac{1}{\pi}d\vec{k}_{//}, \tag{15.26}$$

and Eq. (15.23) becomes:

$$\begin{cases} g_{2D}(E) = \dfrac{S}{2\pi^2}\left(\dfrac{2m^*\pi}{\hbar^2}\right)\displaystyle\sum_n \int_0^\infty \delta\left[E_{//}\left(\vec{k}_{//}\right) + E_{\perp n} - E\right] d\left[E_{//}\left(\vec{k}_{//}\right)\right] \\ g_{2D}(E) = \dfrac{Sm^*}{\pi\hbar^2}\displaystyle\sum_n \int_0^\infty \delta[x + E_{\perp n} - E]dx \end{cases} \tag{15.27}$$

The integral will be zero if the argument of the Dirac function, i.e., $[x + E_{\perp n} - E]$, never reaches zero when the variable x is varied from 0 to $+\infty$. In other words:

$$\begin{cases} \displaystyle\int_0^\infty \delta[x + E_{\perp n} - E]dx = 0 \quad \text{if} \quad [E_{\perp n} - E] > 0 \\ \displaystyle\int_0^\infty \delta[x + E_{\perp n} - E]dx = 1 \quad \text{if} \quad [E_{\perp n} - E] < 0 \end{cases} \tag{15.28}$$

This can be best expressed by considering the step function which is defined as:

$$\begin{cases} \Theta(X) = 0 \quad \text{for} \quad x < 0 \\ \Theta(X) = 1 \quad \text{for} \quad x > 0 \end{cases} \tag{15.29}$$

Therefore, we can write:

$$\int_0^\infty \delta[x + E_{\perp n} - E]dx = \Theta[E - E_{\perp n}], \tag{15.30}$$

and Eq. (15.27) becomes:

$$g_{2D}(E) = \frac{Sm^*}{\pi\hbar^2} \sum_n \Theta[E - E_{\perp n}] \tag{15.31}$$

This relation expresses that, in a quantum well, the density of states of quasi-two-dimensional electrons is a discontinuous function of energy and is incremented by an amount of $Sm^*/\pi\hbar^2$ each time the energy E crosses an allowed value of $E_{\perp n}$, as shown in Fig. 15.7. At each consecutive value of $E_{\perp n}$, a new two-dimensional energy subband begins. The density of states of each new subband is constant so that we obtain the staircase structure shown in Fig. 15.7.

The modification of the density of states in a quantum well (2D) from that in the bulk case (3D), shown in Fig. 15.7, reflects the change in the motion of an electron. The in-plane motion is two-dimensional, which makes the density of states independent on the energy in a subband. For the motion perpendicular to the well plane, we have a new quantum number n, introduced in Eq. (15.19), which replaces one direction of $\vec{k}$ of the three-dimensional case. The excitation of an electron in this direction results in an increase of the quantum number n and thus a transition to the next subband as illustrated by the staircase in Fig. 15.7.

It can be mathematically demonstrated that the density of states for two-dimensional and three-dimensional electrons does coincide at values of $E = E_{\perp n}$, as illustrated in Fig. 15.7, although this is beyond the scope of this discussion.

This considerable dependence of the density of states on the dimensionality of the structure is a key property of low-dimensional structures which opens new possibilities in device applications.

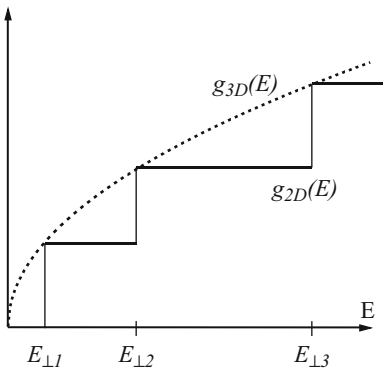


Fig. 15.7 Density of states in the conduction band in a quantum well (2D). The density of states is constant for values of energy between two consecutive quantized energy levels. For comparison, the density of states of a bulk material (3D) is shown in dashed lines

Example
Q: Calculate the number of states between the first and the second energy levels in a quantum well of thickness 25 Å and area of 1 mm^2. Assume that the energy difference between the first two energy levels is 0.3 eV and that the electron effective mass in the quantum well is $m^* = 0.067m_0$ where m_0 is the free electron rest mass.
A: Similar to the three-dimensional case, the number of states is equal to: $N = \int_{E_1}^{E_2} g_{2D}(E)dE$, where E_1 and E_2 are the first and second energy levels in the quantum well, respectively. Since the expression for $g_{2D}(E)$ is given by (we assume $\vec{k}_{//} = \vec{0}$):
$g_{2D}(E) = \frac{Sm^*}{\pi\hbar^2} \sum_n \Theta[E - E_{\perp_n}]$, we obtain:

$$\begin{aligned} N &= \int_{E_1}^{E_2} g_{2D}(E)dE = \frac{Sm^*}{\pi\hbar^2} \sum_n \Theta[E - E_{\perp n}] \\ &= \frac{Sm^*}{\pi\hbar^2}(E_2 - E_1) \\ &= \frac{\left(10^{-3}\right)^2 \left(0.067^* 0.91095 \times 10^{-30}\right)}{\pi\left(1.05458 \times 10^{-34}\right)^2} \left(0.3 \times 1.60218 \times 10^{-19}\right) \\ &\approx 8.40 \times 10^{10} \end{aligned}$$

15.6.3 The Influence of an Effective Mass

In the previous discussion, we have only considered one value for the electron mass m^* for the sake of simplicity. In reality, two effective masses must be considered for the electron in each of the crystals depicted in Fig. 15.5. The effective mass of the electron traveling across the structure thus depends on position, $m^*(x)$. Two Schrödinger equations must then be considered:

$$\begin{cases} -\dfrac{\hbar^2}{2m_1{}^*}\nabla^2\Psi(x,y,z) - [E - U(x,y,z)]\Psi(x,y,z) = 0 \\ \qquad \text{for } x < 0 \text{ and } x > a \\ -\dfrac{\hbar^2}{2m_2{}^*}\nabla^2\Psi(x,y,z) - [E - U(x,y,z)]\Psi(x,y,z) = 0 \\ \qquad \text{for } 0 < x < a \end{cases} \tag{15.32}$$

The other important change concerns the boundary conditions outlined in Eq. (4.52). The continuity of the first derivative of the wavefunction $\partial\Psi(x,y,z)/\partial x$ is no longer valid but must be replaced by the continuity of the product $(1/m^*(x))(\partial\Psi(x,y,z)/\partial x)$, which takes into account the spatial dependence of the electron effective mass. As a result, the boundary conditions in Eq. (4.52) must be replaced by:

$$\begin{cases} \dfrac{1}{m_1^*}\dfrac{\partial \Psi_-}{\partial x}(0) = \dfrac{1}{m_2^*}\dfrac{\partial \Psi_0}{\partial x}(0) \\ \text{and} \\ \dfrac{1}{m_2^*}\dfrac{\partial \Psi_0}{\partial x}(a) = \dfrac{1}{m_1^*}\dfrac{\partial \Psi_+}{\partial x}(a) \end{cases} \tag{15.33}$$

15.7 One-Dimensional Structures: Quantum Wires

15.7.1 Density of States

A quantum wire is formed when the motion of electrons in the conduction band is confined in two directions (e.g., x and y), while it remains free to move in the remaining direction (z). This can be physically achieved by surrounding a small cross-section, rectangular semiconductor crystal with two crystals which have higher bandgap energies.

One way to mathematically treat this situation is to start from the results of a quantum well where the confinement in the x-direction has already been considered and to introduce the confinement in one of the remaining directions (e.g., y). This is not the only way to model quantum wires, and it does not lead to generalized expressions of wavefunctions and energies, but it gives an idea of what is happening. The results can be readily transposed from those of a quantum well and are as follows.

The total wavefunction can be considered as the product of three components:

$$\Psi_{\text{total}}(x, y, z) = \Psi_z(z)\Psi_y(y)\Psi_x(x) \tag{15.34}$$

Only the wavefunction in the z-direction can be easily expressed as a plane wave:

$$\Psi_z(z) = A\exp(ik_z.z) \tag{15.35}$$

where A is a normalization constant. The total energy is the sum of three components:

$$\begin{aligned} E(k_z, n, m) &= E_z(k_z) + (E_x)_n + \left(E_y\right)_m \\ &= \frac{\hbar^2 k_z^2}{2m^*} + \frac{\hbar^2 (k_x)_n^2}{2m^*} + \frac{\hbar^2 \left(k_y\right)_m^2}{2m^*} \end{aligned} \tag{15.36}$$

where n and m are integers (1,2,. . .) used to index the quantized energy levels, $(E_x)_n$ and $(E_y)_m$, and quantized wavenumbers, $(k_x)_n$ and $(k_y)_m$, which result from the confinement of the electron motion in the x- and y-directions, respectively. The values for $(E_x)_n$ and $(E_y)_m$ can be determined, for example, by solving the finite potential well problem in Sect. 4.4.4.

The most important characteristic of a quantum wire is its electron density of states in the conduction band which is given by:

$$g_{1D}(E) = 2\sum_{n,m,k_z} \delta\left[E_z(k_z) + (E_x)_n + \left(E_y\right)_m - E\right] \tag{15.37}$$

In this one-dimensional case, we can make use of the quasi-continuous nature of k_z to write the identity for an arbitrary function $Y(k_z)$:

$$\sum_{k_z} \mathrm{Y}(k_z) \equiv \frac{L}{2\pi}\int_{-\infty}^{+\infty} \mathrm{Y}(k_z)dk_z \tag{15.38}$$

which allows us to simplify Eq. (15.) into:

$$g_{1D}(E) = \frac{2L}{2\pi}\sum_{n,m}\int_{-\infty}^{+\infty} \delta\left[E_z(k_z) + (E_x)_n + \left(E_y\right)_m - E\right]dk_z \tag{15.39}$$

where L is the length of the quantum wire. Moreover, in the one-dimensional case, we have:

$$d[E_z(k_z)] = \frac{\hbar^2}{m^*}k_z dk_z = \frac{\hbar^2}{m^*}\sqrt{\frac{2m^* E_z(k_z)}{\hbar^2}}dk_z \tag{15.40}$$

Therefore, Eq. (15.39) becomes:

$$\begin{cases} g_{1D}(E) = \dfrac{L}{\pi}\sqrt{\dfrac{m^*}{2\hbar^2}}\displaystyle\sum_{n,m}\int_0^{+\infty} \delta\left[E_z(k_z) + (E_x)_n + \left(E_y\right)_m - E\right]\frac{1}{\sqrt{E_z(k_z)}}dE_z \\ \text{or} \\ g_{1D}(E) = \dfrac{L\sqrt{m^*}}{\hbar\pi\sqrt{2}}\displaystyle\sum_{n,m}\int_0^{\infty}\delta\left[x + (E_x)_n + \left(E_y\right)_m - E\right]\frac{1}{\sqrt{x}}dx \end{cases} \tag{15.41}$$

Using Eq. (5.43) and the same argument as Eq. (15.), we obtain:

$$g_{1D}(E) = \frac{L}{\pi\hbar}\sqrt{\frac{m^*}{2}}\sum_{m,n}\frac{\Theta\left(E - \left[(E_x)_n + \left(E_y\right)_m\right]\right)}{\sqrt{E - \left[(E_x)_n + \left(E_y\right)_m\right]}} \tag{15.42}$$

This expression means that, in a quantum wire, the density of states depends on the energy like $1/\sqrt{E}$ in each of the subband defined by two consecutive energy levels $(E_x)_n + (E_y)_m$, as shown in Fig. 15.8.

Equation (15.42) also reveals infinite divergences at points where the energy E coincides with the bottoms of quasi-one-dimensional subbands at $(E_x)_n + (E_y)_m$.

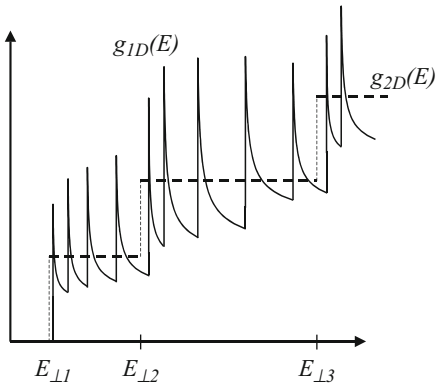


Fig. 15.8 Density of states in the conduction band for a quantum wire (1D). For comparison, the density of states of a quantum well (2D) is shown in dashed lines

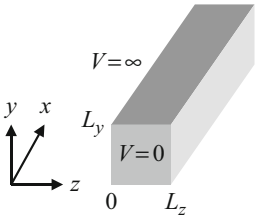


Fig. 15.9 The infinitely deep rectangular cross-sectional quantum wire

These discontinuities take place in an idealized model. In real structures, they are smeared out by the electron collisions mentioned earlier, in Sect. 15.6.1. The maximum values of g_{1D} in Fig. 15.8 are not infinite but correspond to the value of Eq. (15.42) when the denominator is equal to $E - [(E_x)_n + (E_y)_m] \approx \hbar/\tau$, where τ is the electron lifetime discussed earlier.

15.7.2 Infinitely Deep Rectangular Wires

The simplest quantum-wire geometry would have a rectangular cross section surrounded by infinite barriers. This is illustrated schematically in Fig. 15.9 and can be considered to be the two-dimensional analogy to the one-dimensional confinement potential of the standard infinitely deep quantum well.

Within the quantum wires, the potential is zero, while outside the wire, it is infinite. Thus the wavefunction outside the quantum wire should be zero. The form of the potential is $V(y, z) = V(y) + V(z)$, and it is separable. Hence the Schrödinger equation within the wires for the motion along the two directions of confinement (y and z) is:

$$-\frac{\hbar^2}{2m^*}\left[\frac{\partial^2\Psi(y,z)}{\partial y^2}+\frac{\partial^2\Psi(y,z)}{\partial z^2}\right]=E_{y,z}\Psi(y,z) \tag{15.43}$$

The separation of the coordinates in the Schrödinger equation allows the motion to be decoupled further and leads to:

$$\Psi(y,z)=\Psi(y)\Psi(z), \tag{15.44}$$

and then the Schrödinger equation can be written as:

$$-\frac{\hbar^2}{2m^*}\Psi(z)\frac{\partial^2\Psi(y)}{\partial y^2}-\frac{\hbar^2}{2m^*}\Psi(y)\frac{\partial^2\Psi(z)}{\partial z^2}=\left(E_y+E_z\right)\Psi(y)\Psi(z) \tag{15.45}$$

Here the energy components can also be separated into $E_{y,\,z}=E_y+E_z$.
The decoupling is completed with the following equations:

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2\Psi(y)}{\partial y^2}=E_y\Psi(y) \tag{15.46}$$

$$-\frac{\hbar^2}{2m^*}\frac{\partial^2\Psi(z)}{\partial z^2}=E_z\Psi(z) \tag{15.47}$$

The above equations are exactly the same as the infinite quantum well problems (see Sect. 4.4.3, Eq. (4.44) and Eq. (4.45)). The wavefunction solutions are:

$$\Psi(y)=\sqrt{\frac{2}{L_y}}\sin\left(\frac{\pi n_y y}{L_y}\right) \tag{15.48}$$

and

$$\Psi(z)=\sqrt{\frac{2}{L_z}}\sin\left(\frac{\pi n_z z}{L_z}\right) \tag{15.49}$$

which give the components of the energy as:

$$E_y=\frac{\hbar^2\pi^2 n_y^2}{2m^*L_y^2} \tag{15.50}$$

$$E_z=\frac{\hbar^2\pi^2 n_z^2}{2m^*L_z^2} \tag{15.51}$$

Thus, the total energy of the particle due to the confinement is given by the sum of the two discrete components:

$$E_{y,z} = \frac{\hbar^2 \pi^2}{2m^*}\left(\frac{n_y^2}{L_y^2} + \frac{n_z^2}{L_z^2}\right) \tag{15.52}$$

The confined states of a quantum wire are described by the two principal quantum numbers n_y and n_z, and this is in contrast to the single number required for the one-dimensional case discussed in Chap. 4.

15.8 Zero-Dimensional Structures: Quantum Dots

15.8.1 Density of States

An ideal quantum dot, also known as a quantum box, is a structure capable of confining electrons in all three dimensions, thus allowing zero dimension (0D) in their degrees of freedom. In quantum dots, there is thus no possibility for free particle-like motion. The energy spectrum is completely discrete, similar to that in an atom, as will be briefly derived below.

In a quantum dot of rectangular shape, the wavefunction of an electron does not involve any plane wave component, in contrast to other low-dimensional quantum structures. The total energy is the sum of three discrete components:

$$\begin{aligned} E(n,m,l) &= (E_x)_n + \left(E_y\right)_m + (E_z)_l \\ &= \frac{\hbar^2 (k_x)_n^2}{2m^*} + \frac{\hbar^2 \left(k_y\right)_m^2}{2m^*} + \frac{\hbar^2 (k_z)_l^2}{2m^*} \end{aligned} \tag{15.53}$$

where n, m, and l are integers (1,2,...) used to index the quantized energy levels, $(E_x)_n$, $(E_y)_m$, $(E_z)_l$, and quantized wavenumbers, $(k_x)_n$, $(k_y)_m$, and $(k_z)_l$, which result from the confinement of the electron motion in the x-, y-, and z-directions, respectively. The values for $(E_x)_n$, $(E_y)_m$, and $(E_z)_l$ can be determined, for example, by solving the finite potential well problem in Sect. 4.4.4 in all three directions.

As for the quantum wire, the most important characteristic of a quantum dot is its electron density of states in the conduction band which is given by:

$$\begin{aligned} g_{0D}(E) &= 2\sum_{n,m,l} \delta[E(n,m,l) - E] \\ &= 2\sum_{n,m,l} \delta\left[(E_x)_n + \left(E_y\right)_m + (E_z)_l - E\right] \end{aligned} \tag{15.54}$$

There is no further simplification of this expression. The density of states of zero-dimensional electrons consists of Dirac functions, occurring at the discrete energy levels $E(n,m,l)$, as shown in Fig. 15.10.

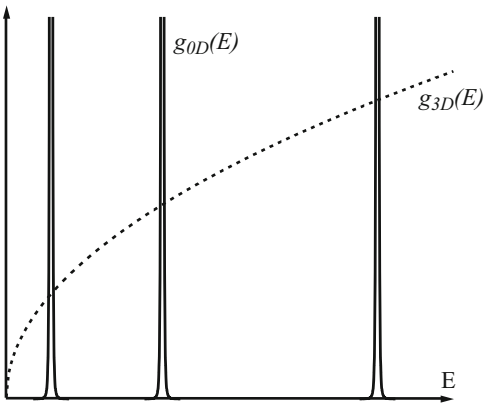


Fig. 15.10 Density of states in the conduction band for a quantum dot (0D). For comparison, the density of states of a bulk crystal (3D) is shown

Again, the divergences in the density of states shown in Fig. 15.10 are for ideal electrons in a quantum dot and are smeared out in reality by a finite electron lifetime τ.

Since quantum dots have a discrete, atom-like energy spectrum, they can be visualized and described as "artificial atoms." This discreteness is expected to render the carrier dynamics very different from that in higher-dimensional structures where the density of states is continuous over a range of values of energy. For example, since all energy states are not allowed, changes in the electron configurations are more restricted.

15.8.2 Infinite Spherical Quantum Dot

The similarity between quantum dots and isolated atoms is close when considering the case of spherical quantum dots, i.e., when the confining potential has a spherical symmetry. For example, nanocrystals in semiconductor-doped glasses and colloidal solutions often have a spherical shape. When the passivation of the surface is made in such a way that carriers are strongly confined in the nanocrystal, the system is usually correctly described by an infinitely deep spherical well, where the confining potential is zero inside and infinite outside a spherical quantum dot with the radius R. The potential can therefore be expressed as:

$$\begin{cases} V\left(\vec{r}\right) = 0 & \text{if } r < R \\ V\left(\vec{r}\right) = \infty & \text{otherwise} \end{cases} \tag{15.55}$$

Due to the spherical symmetry of the potential, the Schrödinger-like equation for the envelope function $\Psi\left(\vec{r}\right)$ in spherical coordinates is given as:

Table 15.2 Values of α_{nl} for the lowest states in a spherical well

n,l	Level	α_{nl}
10	1S	3.142
11	1P	4.493
12	1D	5.763
20	2S	6.283
13	1F	6.988
21	2P	7.725

$$\left[-\frac{\hbar^2}{2m^*}\left(\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)-\frac{\vec{L}^2}{r^2}\right)+V(\vec{r})\right]\Psi(\vec{r})=E\Psi(\vec{r}) \tag{15.56}$$

where $\vec{L}^2$ is the orbital momentum operator which commutes with the Hamiltonian. The solution to Eq. (15.56) is the extension of the one-dimensional problem to the three-dimensional one. The eigenstates are products of the spherical harmonics Y_{lm} and of radial parts given below. The energies and wavefunctions of an infinite spherical quantum dot are:

$$\begin{aligned} E_{nl} &= \frac{\hbar^2}{2m^*}\left(\frac{\alpha_{nl}}{R}\right)^2, \quad n=1,2,3\ldots, \quad l=0,1,2\ldots \\ \Psi(r,\theta,\varphi) &= Aj_l\left(\frac{\alpha_{\mathrm{nl}}r}{R}\right)Y_{lm}(\theta,\varphi) \end{aligned} \tag{15.57}$$

where A is a constant, j_l is a spherical Bessel function, n is the positive integer, and l is the angular momentum quantum number. The coefficients α_{nl} are the zeros of the spherical Bessel functions labeled by an integer in order of increasing energy. Some values of α_{nl} are given in Table 15.2 for the lowest levels defined by n and l. The levels can be labeled with the usual atomic notation, e.g., 1s corresponds to $l=0$ and $n=1$. Their degeneracy is however not the same as in real atoms, and there is no restriction on the values of l for a given n like in free atoms where $l<n$. This is due to the different nature of the potential which in this case encapsulates the particle and its orbit. The degeneracy is only in terms of the allowed "m" values which range from $+l$ to $-l$.

15.9 Optical Properties of Low-Dimensional Structures

Figure 15.11 illustrates the band diagram in a GaAs-AlGaAs quantum well with several electron and hole subbands and the notations used in this section.

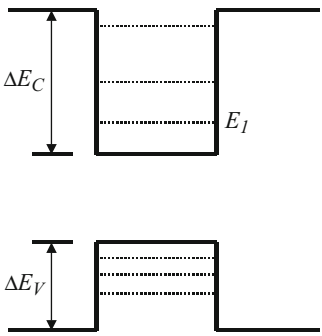


Fig. 15.11 Schematic of band diagram of GaAs-AlGaAs quantum well with electron and hole subbands

15.9.1 Interband Absorption Coefficients of Quantum Wells

The absorption coefficient for a transition from a valence band state of energy E_1 to a conduction band state of energy E_1 has been given earlier in Chap. 10 and can be written as:

$$\alpha(\hbar\omega) = \frac{\pi q^2}{cm_0^2 \bar{n}\varepsilon_0 \omega V} \sum_{1,2} |p_{12}|^2 (f_1 - f_2)\delta(E_{12} - \hbar\omega) \tag{15.58}$$

where:

$$p_{12} = \left\langle 1 \middle| \exp\left(-i\vec{k}_{\lambda} \cdot \vec{r}\right) \vec{e}_{\lambda} \cdot \vec{p} \middle| 2 \right\rangle \tag{15.59}$$

and $\bar{n}$ is refractive index of the medium and V is the volume. f_1 and f_2 are the Fermi occupational probabilities for electrons in the respective states. We assume the incoming photon with the wavevector $\vec{k}_{\lambda}$ and the polarization vector $\vec{e}_{\lambda}$. In the present situation, an electron in the mth heavy-hole subband with 2D wavevector $\vec{k}_{\mathrm{h}}$ absorbs a photon and enters a state with wavevector $\vec{k}_{\mathrm{e}}$ in the nth subband of the conduction band. In terms of the 2D vector $\vec{\rho}$ and the coordinate z normal to the quantum-well layer plane, the wavefunctions are then written as:

$$\begin{aligned} |1\rangle &= \left| h, m, \vec{k}_h \right\rangle = U_h(\vec{\rho}, z) \exp\left(i\vec{k}_h \cdot \vec{\rho}\right) \phi_{hm}(z), \\ |2\rangle &= \left| c, n, \vec{k}_e \right\rangle = U_c(\vec{\rho}, z) \exp\left(i\vec{k}_e \cdot \vec{\rho}\right) \phi_{cn}(z), \end{aligned} \tag{15.60}$$

where U_h and U_c are the cell-periodic parts of the Bloch function and ϕ's are envelope functions. We decompose the photon wavevector as $\vec{k}_\lambda = \left(\vec{k}_{\lambda//}, k_{\lambda z}\right)$ and write Eq. (15.29) as:

$$p_{12} = \left| \left\langle c, n, \vec{k}_e \middle| \exp\left(i\vec{k}_{\lambda//} \cdot \vec{\rho} + ik_{\lambda z} z\right) \vec{e}_\lambda \cdot \vec{p} \middle| h, m, \vec{k}_h \right\rangle \right| \tag{15.61}$$

The matrix element can be evaluated by using Eq. (15.60) for the wavefunctions and integrating over ρ and z. The photon wavevector is considered negligible in comparison to the carrier wavevectors. Thus electron momentum is conserved for the in-plane motion only. However, since the motion is quantized along z-direction, there is no such selection rule for this direction. Using the k-conservation rule and the relation $\vec{k}_\mathrm{e} = \vec{k}_\mathrm{h} + \vec{k}_{\lambda//} \approx \vec{k}_\mathrm{h}$, the squared matrix element can be written as:

$$|p_{12}|^2 = \left\langle |p_{cv}|^2 \right\rangle_{QW} \delta_{\vec{k}_e, \vec{k}_h} C_{mn} \tag{15.62}$$

with

$$C_\mathrm{mn} = |\langle \phi_\mathrm{hm} | \phi_{cn} \rangle|^2 = \left| \int \phi^*_\mathrm{hm} \phi_\mathrm{cn} dz \right|^2 \tag{15.63}$$

In the present case, $\langle |p_{cv}|^2 \rangle_{QW}$ is the polarization-dependent momentum matrix element for transitions between conduction and valence subbands in a quantum well. It is different from the momentum matrix element in bulk semiconductors. The factor $\langle \phi_\mathrm{hm} | \phi_\mathrm{cn} \rangle$ denotes the overlap between the electron and the hole envelope wavefunctions. For infinite potential barriers with parabolic band model, both ϕ_hm and ϕ_cn are sinusoidal functions, and the overlap integral becomes zero unless n is equal to m. Thus in this ideal situation, the optical selection rule is expressed as $C_\mathrm{mn} = \delta_\mathrm{mn}$. However, in real situation the finiteness of the barriers ΔE_c and ΔE_v and also the change in the effective masses of the barriers cause a deviation from the above perfect selection rule.

We can write for the absorption coefficient:

$$\alpha(\hbar\omega) = \frac{\pi q^2}{c m_0^2 \bar{n} \varepsilon_0 \omega V} C_{mn} \sum_{\vec{k}_\mathrm{e}, \vec{k}_\mathrm{h}} \left\langle |p_{cv}|^2 \right\rangle_{QW} \delta_{\vec{k}_\mathrm{e}, \vec{k}_\mathrm{h}} (f_\mathrm{e} - f_\mathrm{h}) \delta\left(E_\mathrm{e}\left(\vec{k}_\mathrm{e}\right) - E_\mathrm{h}\left(\vec{k}_\mathrm{h}\right) - \hbar\omega \right) \tag{15.64}$$

Using the parabolic $E\left(\vec{k}\right)$ relation, the energies are expressed as:

$$\begin{cases} E_{\mathrm{e}}\left(\vec{k}_{\mathrm{e}}\right) = E_{\mathrm{cn}} + \dfrac{\hbar^2 k_{\mathrm{e}}^2}{2m_e} \\ E_{\mathrm{h}}\left(\vec{k}_{\mathrm{h}}\right) = -E_{\mathrm{g}} - E_{\mathrm{hm}} - \dfrac{\hbar^2 k_{\mathrm{h}}^2}{2m_{\mathrm{h}}} \end{cases} \tag{15.65}$$

The Fermi occupational probability can be written as:

$$f(E) = \frac{1}{1 + \exp\left[\left(E - E_f\right)/k_{\mathrm{b}}T\right]} \tag{15.66}$$

where E_{f} is the quasi-Fermi level. Using $\vec{k}$-conservation, the double summation in Eq. (15.64) is reduced to a single summation over $\vec{k}_{\mathrm{h}}$. The argument of the energy-conserving δ-function then becomes:

$$E_{\mathrm{e}}\left(\vec{k}_{\mathrm{e}}\right) - E_{\mathrm{h}}\left(\vec{k}_{\mathrm{h}}\right) - \hbar\omega = \left(E_{\mathrm{g}} + E_{\mathrm{cn}} + E_{\mathrm{hm}}\right) - \frac{\hbar^2 k_{\mathrm{h}}^2}{2m_r} - \hbar\omega, \tag{15.67}$$

where m_r is the reduced mass (Eq. 10.80). The remaining sum in Eq. (15.64) becomes:

$$\sum_{\vec{k}_{\mathrm{h}}} \rightarrow \frac{2S}{(2\pi)^2} \int \delta\left(E_{\mathrm{g}} + E_{\mathrm{cn}} + E_{\mathrm{hm}} - \frac{\hbar^2 k_{\mathrm{h}}^2}{2m_{\mathrm{r}}} - \hbar\omega\right) 2\pi k_{\mathrm{h}} dk_{\mathrm{h}} \left[f_{\mathrm{h}}(k_{\mathrm{h}}) - f_{\mathrm{e}}(k_{\mathrm{h}})\right] \tag{15.68}$$

where S is the area of the quantum well and factor 2 is from spin degeneracy. The integration in Eq. (15.68) is performed easily due to the presence of the δ-function, so that we obtain:

$$\alpha(\hbar\omega) = \frac{m_r q^2 C_{\mathrm{mn}} \left\langle |p_{cv}|^2 \right\rangle_{QW}}{\varepsilon_0 \hbar^2 c m_0^2 \bar{n} \omega L} \left(f_{\mathrm{e}} - f_{\mathrm{h}}\right) H\left(\hbar\omega - E_{\mathrm{g}} - E_{\mathrm{cn}} - E_{\mathrm{hm}}\right) \tag{15.69}$$

where L is the thickness of the quantum well and $H(x)$ is the Heaviside step function. Equation (15.69) may be compared with the expression for bulk. Remember from Chap. 5 that $|p_{vc}|^2$ can be expressed and estimated in terms of the Kane matrix elements (Eq. (5.67) and Sect. 5.7). In both cases, the absorption coefficients are proportional to the respective joint density of states function. The expected variation of absorption coefficients are shown in Fig. 15.12. The experimental measurements of absorption coefficient in GaAs/AlGaAs quantum wells and thick GaAs layer are compared in Fig. 15.12.

When considering intersubband absorption, we immediately have the selection rule that normal incident light with (x,y) polarization cannot be absorbed because the z-confined wavefunctions are orthogonal. In order for light to be absorbed in an intersubband transition, it is essential that there should also be a z-polarized component giving an qzE_z coupling term.

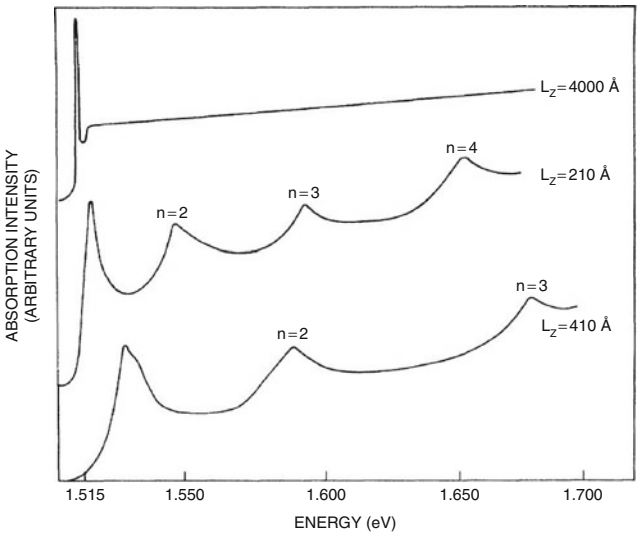


Fig. 15.12 Absorption coefficient in GaAs/AlGaAs quantum wells and thick GaAs layers (upper curve). The peaks correspond to quantum-confined subband n (Reprinted figure with permission from Dingle R, Wiegmann W, and Henry CH, Phys Rev Lett (Vol. 33): p. 829, Fig. 2. Copyright 1974 by the American Physical Society)

15.9.2 Absorption Coefficient of Quantum Wires

The calculation of the absorption coefficient may be performed as usual by assuming the $\vec{k}$-conservation condition to be valid along the direction of the free motion. The absorption coefficient is written as:

$$\alpha(\hbar\omega) = -B_1 \sum_{\vec{k}} (f_e - f_h)\delta\left(E_g + \hbar^2 k^2/2m_r - \hbar\omega\right) \tag{15.70}$$

where B_1 is a constant and E_g denotes the effective gap which is bulk bandgap plus the electron and hole subband energies. The summation over $\vec{k}$ may be converted into an integral, and assuming $f_e - f_h = 1$ the integration may be performed to yield:

$$\alpha(\hbar\omega, a) = -\frac{q^2 C_{1D}\left\langle |p_{cv}|^2 \right\rangle_{QWR} (2m_r)^{1/2}}{2m_0^2 \varepsilon_0 \bar{n}\omega c S} \left(\hbar\omega - E_g\right)^{-1/2} \tag{15.71}$$

where the coefficient C_{1D} is the overlap integral of a quantum wire and $\langle |p_{cv}|^2 \rangle_{QWR}$is the momentum matrix element for transitions between conduction and valence subbands in a quantum wire. S is the cross-sectional area of the wire.

Equation (15.71) leads to the conclusion as noted before that the absorption coefficient is proportional to the joint density of states function. Therefore the absorption coefficient should show a singularity at $\hbar\omega = E_g$ and fall with increasing photon energy as shown in Fig. 15.8.

15.9.3 Absorption Coefficient of Quantum Dots

The absorption coefficient of a cubic QD system of side length a may be written as:

$$\alpha(\hbar\omega) = \frac{2\pi q^2 \left\langle |p_{cv}|^2 \right\rangle}{m_0^2 \bar{n} \varepsilon_0 c \omega a^3} \sum_m g(m^2) \delta\left(\hbar\omega - E_g - \frac{\pi^2 \hbar^2 m^2}{2 m_r a^2}\right) \tag{15.72}$$

where $g(m^2)$ is the degeneracy of the energy level determined by m^2. Only $\Delta m = 0$ transitions are allowed. Equation (15.72) indicates that the interband absorption in a QD will be a series of discrete lines, representing the reduced density of states function of a 0D system. The discrete lines will occur at photon energies:

$$\hbar\omega = E_g + \frac{\pi^2 \hbar^2 m^2}{2 m_r a^2} \tag{15.73}$$

In practice the absorption spectra are not discrete lines but are broadened because of the size distribution of quantum dots. We consider that the family of dots has a fluctuation in side length described by the following Gaussian distribution:

$$P(a) = \frac{1}{(2\pi)^{1/2} D} \exp\left[-\frac{(a - a_0)^2}{2D^2}\right] \tag{15.74}$$

where a_0 is the average value and $D^2 = \langle (a - a_0)^2 \rangle$ is the standard deviation.

Using Eq. (15.72) and Eq. (15.74), the absorption coefficient for a nonuniform quantum dot system can be calculated as:

$$\alpha = \int_0^\infty P(a) \alpha(\hbar\omega, a) da \tag{15.75}$$

The line broadening also occurs due to phonon scattering processes in addition to the size distribution of QDs.

15.10 Examples of Low-Dimensional Structures

The optical properties of low-dimensional quantum structures, arising from their peculiar density of states, are often put to use in semiconductor optoelectronic devices, such as semiconductor laser diodes and quantum-dot infrared photodetectors. Such low-dimensional structures are fabricated in practice using a succession of processes involving epitaxy, lithography, and etching. An illustration of the principle of quantum wells, wires, and dots is shown in Fig. 15.13.

Low-dimensional quantum structures have, for example, been most beneficial for semiconductor laser diodes, leading to low threshold current (minimum necessary current for lasing), high power, and weak temperature-dependent devices. These

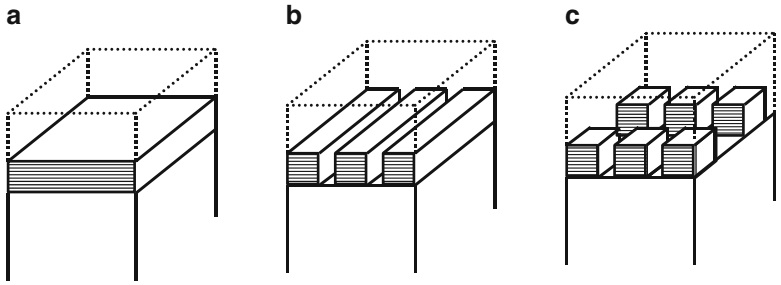


Fig. 15.13 Illustration of a (**a**) 2D structure (quantum well), (**b**) 1D structure (quantum wires), and (**c**) 0D structure (quantum dots), showing the various levels of spatial confinement

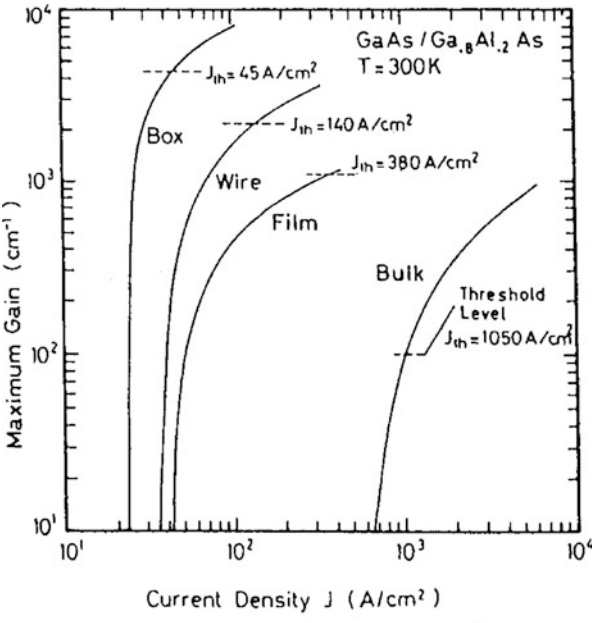


Fig. 15.14 Coefficient of light amplification (gain) for different structures. The dashed lines show the threshold current density above which laser emission starts (Reprinted with permission from IEEE Journal of Quantum Electronics Vol. 22, Asada M, Miyamoto Y, and Suematsu Y, "Gain and the threshold of 3-dimensional quantum-box lasers," p. 1918, Fig. 6. Copyright 1986, IEEE)

properties, in conjunction with their small size, have made laser diodes attractive for applications involving densely packed laser arrays. This applies also to the monolithic integration of lasers with low-power electronics such as computer optical interconnects, optoelectronic signal processing, and optical computing.

An illustration of the effect of low-dimensional quantum structures on the properties of optoelectronic devices is shown in Fig. 15.14 which illustrates the theoretical predictions for threshold currents in semiconductor lasers based on active regions with different low-dimensional structures. By using quantum dots instead of a bulk layer, the threshold current may be reduced by more than 20 times. This is due to the abrupt energy dependence of the density of states in low-dimensional quantum structures which can enhance the light amplification mechanisms and thus allows lasing to occur at lower currents.

15.10.1 Quantum Wires

Figure 15.15 shows an example of a quantum wire, which has been etched in a thin film of doped GaAs deposited on an undoped GaAs substrate. Inside the rectangular stripe, there is a highly conductive channel where the electrons are confined and which forms a quantum wire and whose width is narrower than that of the stripe. In GaAs wires, the minimum diameter of the channels can be about 80 nm.

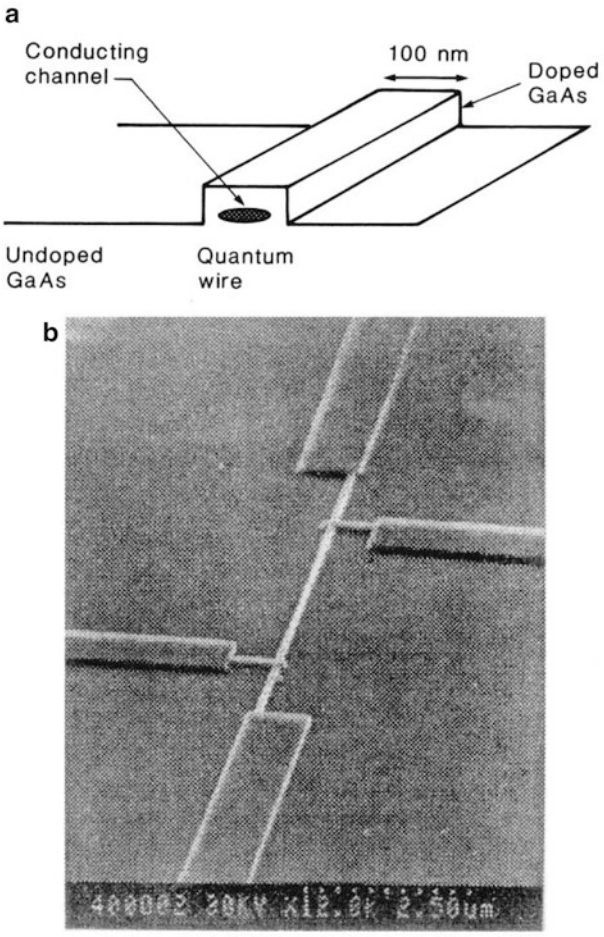


Fig. 15.15 Quantum wire formed by etching away all but a thin strip of doped semiconductor on an undoped substrate: (**a**) schematic diagram; (**b**) practical example ("Fig. 11.1", from Low-Dimensional Semiconductors: Materials, Physics, Technology, and Devices by M.J. Kelly; taken after Physica Scripta Vol. T45, Beaumont SP, "Quantum wires and dots: defect related effects," p. 196. Copyright 1992, Physica Scripta. Reprinted with permission of Oxford University Press, Inc. and The Royal Swedish Academy of Sciences)

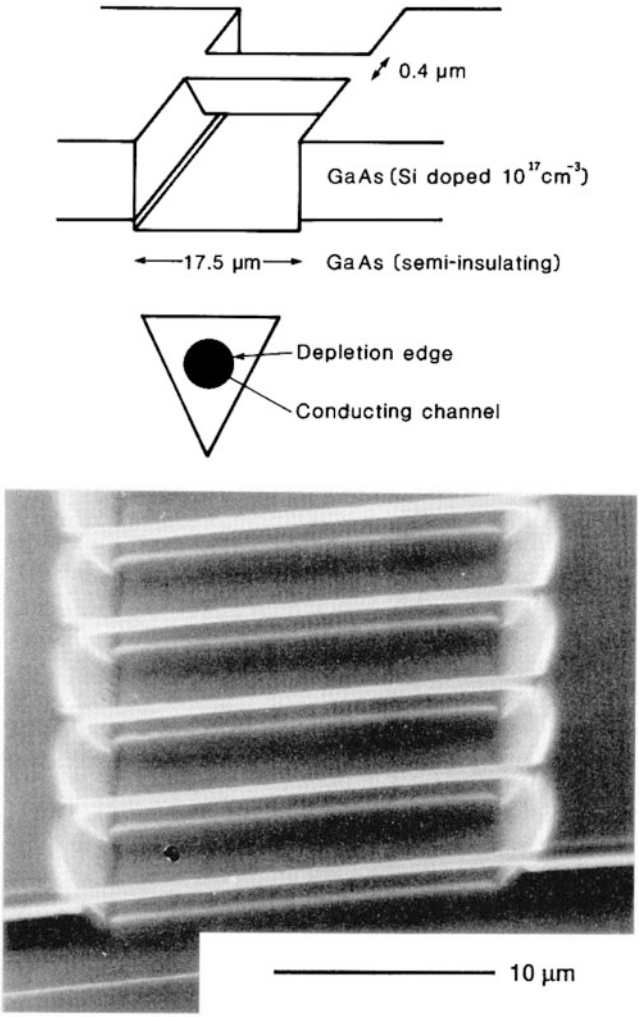


Fig. 15.16 Schematic diagram and image of quantum wires of doped GaAs on an insulating substrate (Taken from Fig. 11.2 of "Low-Dimensional Semiconductors: Materials, Physics, Technology, and Devices" by MJ Kelly; taken after Journal of Vacuum Science and Technology B Vol. 6, Hasko DG, Potts A, Cleaver J.R.A., Smith CG, and Ahmed H, "Fabrication of submicrometer freestanding single-crystal gallium arsenide and silicon structures for quantum transport studies," p. 1851. Copyright 1988, American Institute of Physics. Also taken after Physica Scripta Vol. T54 Kelly, et al." Quasi-One-Dimensional Transport in Semiconductor Microstructures" p.201, Fig. 1 (a), Copyright 1992 Royal Swedish Academy. Reprinted with permission of Oxford University Press, Inc., American Institute of Physics)

Another example of quantum wire is shown in Fig. 15.16. The structure was made by using etching a doped thin GaAs film, in such a way that it undercuts the crystal from the surface, i.e., GaAs material is removed *below* the remaining stripe.

The resulting structure thus has a triangular cross section, and a highly conductive channel is present inside it which is where the electrons are confined and a quantum wire is formed.

Quantum wires have novel optical absorption spectra which depend on the polarization of the light. The optical properties can be computed with the methods we discussed in Chap. 10. Again the key quantity and novelty will be mainly due to the joint density of states. But more recently, scientists have discussed another reason why the quantum wire may be of interest, and this is in the context of electron-electron interactions having a stronger effect on carrier mobility. The dense many-electron quantum wire is also called the Luttinger liquid (Bockrath et al. 1999) and exhibits an exciting new science which has been studied only very recently. When moving along a "line," carriers are more likely to be affected by each other's Coulomb interaction. A carrier will find it difficult or even impossible in some cases to pass another charge or to avoid the other charge, if for some reason this charge is blocked on the way. One trapped carrier in the wire can stop the entire flow of current, which is an example of the Coulomb blockade. The controlled blockage and removal of the blockage is one of the targets of present-day nanotechnology research. In this way, the presence or absence of a single charge in a trap can give rise to a measurable quantity of electrical current. The quantum wire is especially interesting if all the electron spins are pointing in one direction. This can be done either because they have been aligned by a magnetic field or because they have been injected into the wire by a magnet. Quantum wires can therefore be used as "spin wires" which transport spin information from one area of a device to the other.

The fabrication of quantum pillars or vertical quantum wires, as shown in Fig. 15.7, in doped (multilayer) semiconductors is more complicated. The processing steps are shown in Fig. 15.17c and d which result in the structure shown in Fig. 15.17b. A submicrometer-diameter metal dot is laid down onto the film (step 2 in Fig. 15.17c), and the pillar structures are formed by an etching process (step 3), through which parts of the material are selectively removed. The electrons are thus confined laterally inside the pillars (4). This structure is then filled with polyimide, a polymeric material, and etched back to expose the top of the metal dot (steps 1 and 2 in Fig. 15.17d). The whole surface can then be coated with metal (steps 3 and 4), making contact to the metal dots and thus the vertical quantum wire. The fabrication methods can be refined so that any single pillar can be contacted.

15.10.2 Quantum Dots

The structures shown in Fig. 15.17 can also be used as a quantum dot if the carriers can be confined vertically at the top and bottom of the pillars, in addition to being confined laterally by the side walls of the pillars. This can be achieved by choosing the two barrier layers (AlGaAs in Fig. 15.17a) that are sufficiently thick.

Another method of realizing semiconductor quantum dots consists of making use of a strain-induced transformation that occurs naturally in the initial stages of growth of lattice-mismatched materials. This type of growth usually starts atomic layer by

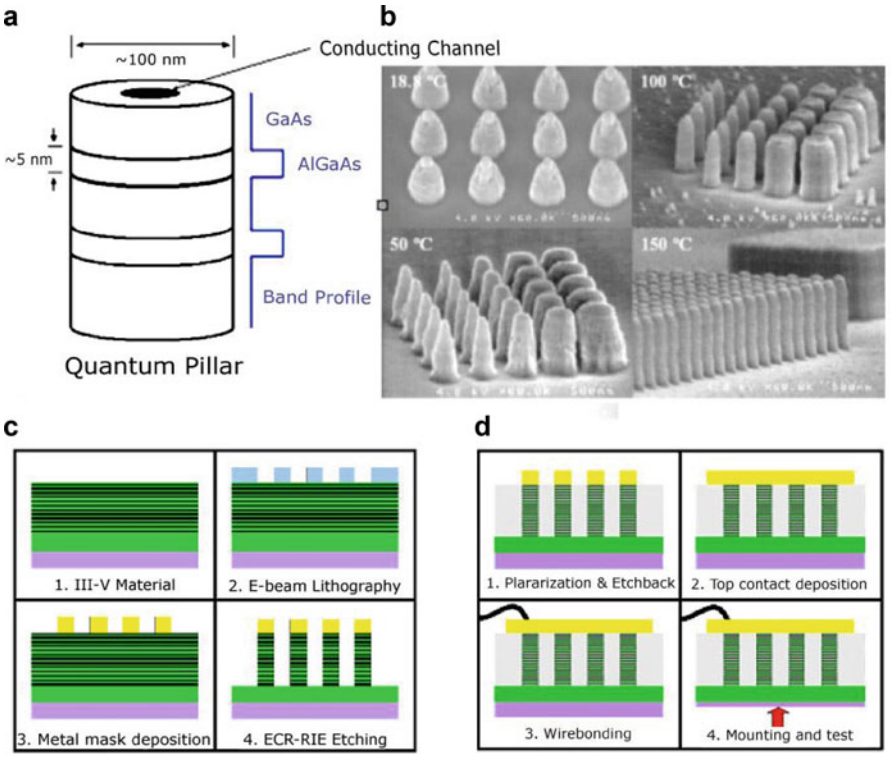


Fig. 15.17 A quantum pillar formed from resonant tunneling semiconductor multilayers showing (**a**) a schematic diagram of the pillar, (**b**) the partially processed structure after the first etch, and (**c**) and (**d**) the full processing route

atomic layer, and after a certain critical thickness is reached, nanometer-size islands spontaneously form. This is known as the Stranski-Krastanov growth mode. These islands show good size uniformity and large surface densities. In this method, the growth has to be interrupted immediately after the island formation and before the islands reach a size for which strain relaxation and defects occur. This spontaneous island formation during growth precludes the interface quality problems often associated with low-dimensional quantum structures achieved through etching. This breakthrough has created some excitement in the physics community by providing the opportunity for experimental verification of the effects of three-dimensional quantum confinement in semiconductor structures.

Several reports worldwide show remarkable agreement on the optical properties of these structures, finding that the delta function density of states expected from 0D quantum structures manifested itself in ultrasharp light emission peaks. Compound semiconductors that have been used until now for quantum dots include InAs and InGaAs on GaAs, InAlAs on AlGaAs, InAs on InP, and InP on InGaP and GaP.

15.10.3 Effect of Electric and Magnetic Fields

In the confined direction of the quantum well or in nanopillars and quantum dots, the electrons subjected to an electric field cannot wonder away to infinity, so the electric field constitutes a relatively small perturbation and can be handled by methods of quantum-mechanical perturbation theory. The same is true for nanopillars and quantum dots in a magnetic field. The expansion of energy levels and wavefunctions can be usually stopped in second order, giving us a powerful way of estimating field-induced changes in energies and optical permittivities. In Chap.10, we showed how permittivity can be related to the wavefunctions and energy spectrum (Eq. 10.53)). In static fields, we can work with the time-independent Schrödinger equation and perturbation theory. Following Sect. 4.3.1, we can calculate the new ground-state wavefunction to first-order and energy to second-order perturbation theory using Eq. (4.205) and Eq. (4.207). The perturbation caused by an electric field E_{0z}applied in the z-direction is $-qzE_{0z}$ giving the second order (first order vanishes by symmetry in a symmetric coordinate system):

$$E_{\mathrm{g}}^{(2)} = \sum_{l} (aE_{0z})^2 \frac{\left|z_{gl}\right|^2}{E_g - E_l} \tag{15.76}$$

In a confined system, the electric field-induced shift of the energy of the free subband eigenstates is called the Stark shift and is lowering of energy when we start with box eigenstates. Figure 15.18 shows the absorption spectra of a quantum well in

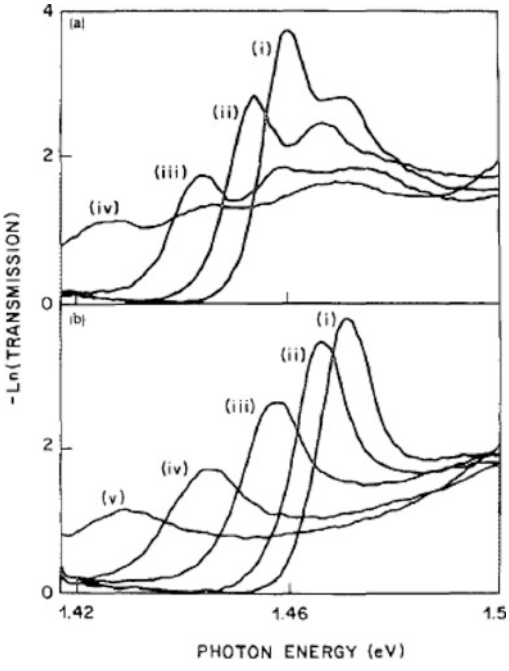


Fig. 15.18 Electro-absorption spectra of GaAs quantum-well waveguide device as a function of electric field applied field perpendicular to the plane of the layers. (i) = 1.6×10^4 V·cm^{-1}; (ii) = 10^5V·cm^{-1}; (iii) = 1.4×10^5V·cm^{-1}; (iv) = 1.8×10^5V·cm^{-1}; (v) = 2.2×10^5V·cm^{-1} (Reprinted with permission from Applied Physics Letters Vol. 47, Weiner JS, Miller D. A.B., Chemla DJ, Damen TC, Burrus CA, Wood T H, Gossard AC, Wiegmann W, "Strong polarization sensitive electro-absorption in GaAs/AlGaAs quantum well waveguides," p. 1149. Copyright 1985, American Institute of Physics)

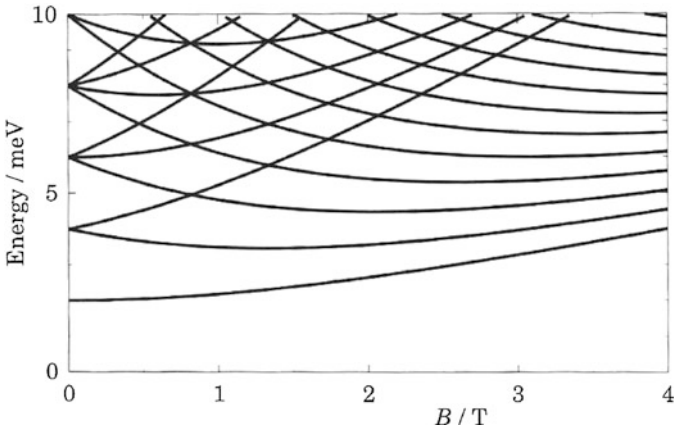


Fig. 15.19 The energy levels of a parabolically confined quantum dot with intrinsic energy level splitting $\hbar\omega_0 = 2$meV in a magnetic field (Davies JH, The Physics of Low Dimensional Semiconductors: an Introduction, p. 237, Fig. 6.16. © Cambridge University Press 1998. Reprinted with the permission of Cambridge University Press)

an electric field applied perpendicular to the layers and also shows the Stark energy shift of the exciton peak. The action of an electric field on an exciton can however in some cases be more complex than just a Stark shift, especially when the exciton is broken up by the field, and then the simple method might not suffice.

Figure 15.18 shows the effect of a magnetic field on the energy levels of a large quantum dot in which electrons are confined by a three-dimensional parabolic potential, with energy levels at ~2 meV interval (Fig. 15.19).

In this example the magnetic energy levels and the intrinsic confinement level splittings are comparable at $B = 1$ T, so the effect of the B field is obviously large. In smaller dots, one needs a correspondingly larger B field to see the same relative shifts or a smaller effective mass. When the magnetic coupling is treated in perturbation, both the first-order and the second-order terms contribute to the energy. In the notation of Chap. 10 and from Eq. (10.113), the perturbation is of the form (m^* is the effective mass):

$$V = \left[(qBx)^2 - 2qBx\left(i\hbar\frac{\partial}{\partial y}\right)\right]\frac{1}{2m^*} \tag{15.77}$$

The first-order perturbation shift in energy is positive, and the second-order term is necessarily negative. The B field will in general raise the energy of the electron when it is in the ground state.

Finally Fig. 15.20 shows the drastic effect a magnetic field has on the longitudinal and Hall resistance of a high-quality, high-mobility quantum well. The magnetic field is applied perpendicular to the plane in which conduction takes place. We explained in Chap. 10 how the magnetic field produces Landau levels and how the degeneracy of the levels changed with B and that the Fermi energy in Landau levels

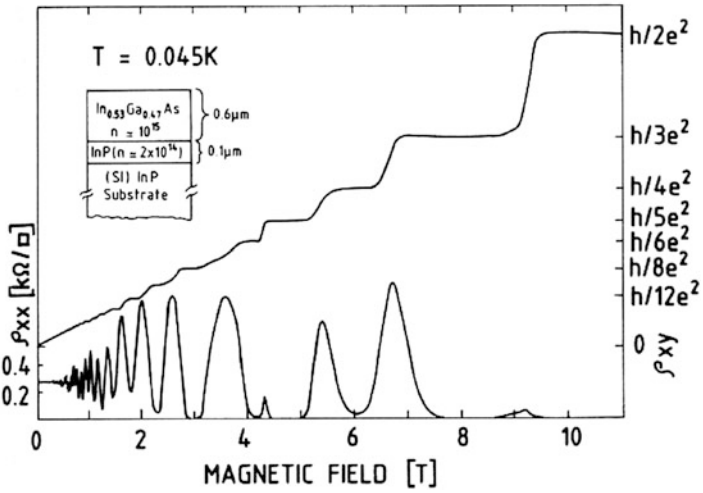


Fig. 15.20 Shubnikov-de Haas trace (ρ_{xx}) and quantum Hall effect (ρ_{xy}) as a function of magnetic field normal to the plane at $T = 0.045$ K in $Ga_{0.47}In_{0.53}As$-InP heterostructures (Reprinted with permission from Applied Physics Letters Vol. 48, Razeghi M, Duchemin JP, Portal JC, Dmowski L, Remeni G, Nicolas RJ, and Briggs A, "First observation of the Quantum Hall effect in a $Ga_{0.47}In_{0.53}As$-InP heterostructure with three electric subbands," p. 712. Copyright 1986, American Institute of Physics)

moves with B field for a given electron concentration. Increasing the magnetic field increases the Landau level splittings and the degeneracy of each band. This then implies that the density of states at the Fermi level changes too. The Fermi level can move from a region of finite to a region of zero density of states, i.e., sit in the gap between two adjacent Landau levels. But this then according to Eq. (10.39) drastically changes the longitudinal resistance with B field exactly as shown in Fig. 15.20. In contrast to the longitudinal resistance, we see that the Hall resistance does not vanish when the Fermi level is in the Landau gaps but forms plateaus until the Fermi level again crosses into the middle of the next Landau band, at which point the resistance suddenly goes up again with B field. This fascinating phenomenon is known as the quantum hall effect or QHE. The plateau signifies that in this interval of level filling (B decreasing) or emptying (B increasing), the number of Hall carriers is not changing. We see a plateau in the gap and not zero conductance because the Hall voltage is not a Fermi level property. When the Fermi level crosses a region of small density of states, i.e., from the maxima through the gaps, then it is passing through energy levels which are spatially localized; the orbits of the localized states form closed paths which do not intersect the sample edge. The energy levels which are affected by the B field are the delocalized ones which sit in a narrow region in the maxima and obey $\varepsilon_n = (n + 1/2)\hbar\omega_c$. Remember that in the semiclassical description, the Hall voltage exists because the Lorentz force creates an asymmetric charge redistribution for drifting carriers.

15.11 Summary

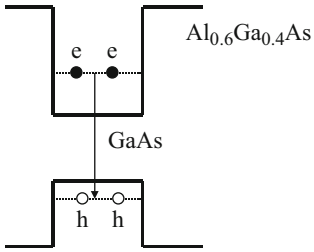
In this chapter, we have first reviewed topics associated with semiconductor heterostructures. In particular, the concepts of type I and type II band alignments were outlined. Furthermore, model solid theory and Anderson's model for heterojunction energy band alignment and diagram were described.

Subsequently, we showed that the motion of electrons in a crystal can be spatially confined in one, two, or even three directions, by designing and fabricating an adequate semiconductor structure, a quantum well, wire, or dot. When the amount of confinement is sufficient, quantum-mechanical effects become important and lead to the discretization of the energy spectrum, i.e., the quantization of allowed energy levels becomes an important feature of the system. A rough criterion is as always $\Delta E_{n,\ n\ +\ 1} \sim k_{\mathrm{b}}T$, i.e., the splitting has to be bigger or comparable to the thermal energy.

The important new characteristic of a low-dimensional quantum structure is the new density of states. This quantity shows a different dependence on energy, especially for lower (wire and dot) dimensionality systems. The magnitude and energy dependence of the density of states strongly correlates with many properties of the solid and in particular with the optical properties of a semiconductor. This has been shown here and in Chap. 10. We have shown how electric and magnetic fields affect confined eigenstates and eigenenergies. Ironically it is often easier to estimate the effect of external fields in confined systems than in infinite ones because energy levels are discrete and wavefunctions normalized in a small volume. This means that one can use standard second-order perturbation theory. Confinement can be exploited in the design of the characteristics of optoelectronic devices. Having evaluated changes to energies and wavefunctions, it is possible to compute the electro-optic coefficients using the methods of Chap. 10 combined with the perturbation expansion given here.

Problems

1. In this chapter, we used the effective mass of the electron in the Schrödinger equation. Explain why it was necessary to do so, whereas it was not necessary in the infinite and finite potential well in Chap. 4.
2. Give an expression (an integral) for the total number of electrons in the conduction band of a bulk three-dimensional semiconductor and then in the first subband of a quantum well of width L in terms of the density of states and the Fermi function (assume box eigenstates in the confined direction). If at $T = 0$ K we dope the first subband in the conduction band and we fill all the states in the first subband, how many electrons do we need per unit area?
3. Consider a 50 Å GaAs and 300 Å $Al_{0.6}Ga_{0.4}As$ layers forming a quantum well structure.



The electrons are all located at the first energy state (e), and holes are at (h). The general expression of the first energy state is determined as $E_1 = \hbar^2\pi^2/2m^*a^2$, where a is the width of the quantum well and m^* is the effective mass of the particle considered (for holes, consider the heavy-hole effective mass).

What is the photon energy of the light emitted when the electron and the hole recombine as shown in the above figure?

4. Let us assume a quantum dot which is spherical. The electrons or holes are confined at energy states with the following expression: $E_{nl} = (\hbar^2/2m^*)(\alpha_{\mathrm{nl}}/R)^2$, where m^* is the effective mass of the electron or hole and the value of α_{nl} is given by $\alpha_{10} = \pi$, $\alpha_{11} = 4.49$, $\alpha_{12} = 5.76$, $\alpha_{20} = 6.28$, $\alpha_{21} = 7.72$, and $\alpha_{22} = 9.09$. Now consider the very small GaAs quantum dots of radius 10 nm. If the electron drops from the second state (α_{11}) to the first state (α_{10}), what is the photon energy of the light emitted from this transition?

   Draw the energy diagram for GaAs quantum dots with radius 5 nm, 10 nm, and 15 nm. How does the first energy state change as a function of radius?

5. *Density of States of an Ideal Two-Dimensional Electron Gas*

   Using the infinite barrier approximation, derive an expression for the density of states for electrons in a quantum well in terms of the well width L and electron effective mass m^*.

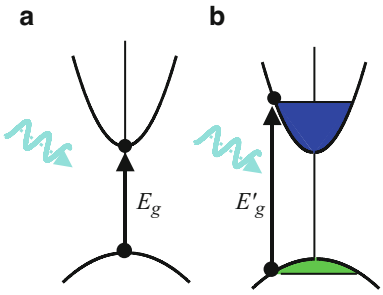
6. *Fermi* Energy *of an Ideal Two-Dimensional Electron Gas*

   Consider a structure consisting of two GaAs quantum wells that have been grown far apart in $Al_xGa_{1-x}As$ with the same Al composition x ($x \leq 0.3$). In well A the GaAs thickness is L, while in well B, it is $2L$. Now, approximate the conduction bands in wells A and B by ideal quantum wells between infinitely high potential barriers. Suppose that the quantum wells contain electrons and that both wells have the same Fermi energy,$E_F = 3F_1^A$ where E_1^A is the lowest quantized energy level in well A.

   (a) How many subbands in each well contain electrons at zero temperature?
   (b) What is the two-dimensional charge density N_A and N_B in each well? Give the answer in terms of known physical quantities such as $\hbar$ and L.

7. *The* Graphic *of the Two-Dimensional Density of States*
   Figure 15 shows the density of states of a quantum well. The confinement energy of the lowest level (E_1) is 17 meV, and the first excited state (E_2) has a confinement energy of 30 meV. The Fermi level is located 50 meV above the bottom of the conduction band. Determine the number of electrons contained in the well.
8. The two-dimensional potential which confines the electrons in a quantum wire made of GaAs is assumed to be parabolic, and the subband separation is given as $\hbar\omega_0 = 12$ meV. If the Fermi energy is $E_F = 37$ meV as measured from the bottom of the lowest subband, calculate the number of electron per unit length at 0 K including spin degeneracy.
9. *The* Moss-*Burstein Shift in Absorption Spectra*
   The "band filling" or Moss-Burstein shift effect occurs in all heavily doped three-dimensional semiconductors. It is a consequence of the fact that electrons are fermions, and therefore it is impossible (by the Pauli exclusion principle) to optically excite an electron into a same spin k-state, which is already occupied. In the case of strongly degenerate n^+-doped sample, this has the effect of prohibiting any interband transition into electron states below the Fermi energy leading to an upward shift in the effective absorption edge, E'_g (see figure below). The *Moss-Burstein shift* (ΔE) is defined as the difference between the effective absorption edge and the energy gap (E_g) of the material, i.e., $\Delta E = E'_g - E_g$.

   (a) Calculate the *Burstein shift* in the absorption edge of a direct semiconductor with parabolic bands due to the heavy doping (n-type) at very low temperature ($T \approx 0$ K). The carrier concentration is n_e. Neglect excitons. Note that the shift is not simply the Fermi energy of the electrons and involves the mass of both conduction and valence bands.
   (b) Calculate the *Moss-Burstein shift* for the GaAs material doped with 1×10^{18} electrons/cm^3. What should happen to the shape of the absorption edge? Assume $m_e^* = 0.067\, m_0$ and $m_h^* = 0.45\, m_0$ where $m_0$is the electron rest mass.



Allowed optical transitions in direct-gap semiconductors: (**a**) undoped material, absorption threshold of E_g; (**b**) n^+-doped material, absorption threshold blue shifted to E_g' by the Moss-Burstein shift

10. *Critical Radius of a Spherical Quantum Dot with Finite Barrier Height*
Assume that a quantum dot has a spherical shape with radius R and is surrounded by a medium of higher bandgap such as AlGaAs. The potential barrier at the conduction band is ΔE_c at all points in the surface of the sphere. The potential well is a square well of height, ΔE_c, for $r > R$ and is 0 for $r < R$. Let us consider the simplest case of zero angular momentum ($l = 0$), and then it follows that the wavefunctions $\Psi(\vec{r})$ depends only on the radial part. When $l = 0$ and $\Psi(\vec{r}) = R(r) = \phi(r)/r$, Eq. (15.56) reduces to:

$$\begin{cases} -\dfrac{\hbar^2}{2m^*}\dfrac{d^2\phi(r)}{dr^2} + \Delta E_c\phi(r) = E\phi(r) & (r > R) \\ -\dfrac{\hbar^2}{2m^*}\dfrac{d^2\phi(r)}{dr^2} = E\phi(r) & (r < R) \end{cases}$$

The solution of the above equation is the same as the one with a one-dimensional finite potential well. Find the critical radius below which there is no bound state of one electron in the quantum dot.

References

Bockrath M, Cobden DH, Lu J, Rinzler AG, Smalley RE, Balents L, McEuen PL (1999) Luttinger liquid behaviour in carbon nanotubes. Lett Nat 397:598–601
Shockley W (1951) U.S. Patent 2,569,347

Further Reading

Ahmed H (1986) An integration microfabrication system for low dimensional structures and devices. In: Kelly MJ, Weisbuch C (eds) The physics and fabrication of microstructures and microdevices. Springer, Berlin, pp 435–442
Ashcroft NW, Mermin ND (1976) Solid state physics. Holt, Rinehart, Winston, New York
Bassani F, Pastori Parravicini G (1975) Electronic states and optical transitions in solids, Chap. 6. Pergamon, New York
Bastard G (1988) Wave mechanics applied to semiconductor Heterostructures. Halsted Press, New York
Beaumont SP (1992) Quantum wires and dots-defect related effects. Phys Scr T45:196–199
Chuang SL (1995) Physics of optoelectronic devices. John Wiley & Sons, Inc., New York
Davies JH (1998) The physics of low dimensional semiconductors: an introduction. Cambridge University Press, New York
Dingle R Confined carrier quantum states in ultrathin semiconductor heterostructures. In: Queisser HJ (ed) Feskorperproblem XV, pp 21–48
Einspruch NG, Frensley WR (1994) Heterostructures and quantum devices. Academic Press, London

Hasko DG, Potts A, Cleaver JR, Smith C, Ahmed H (1988) Fabrication of sub-micrometer free standing single crystal GaAs and Si structures for quantum transport studies. *J Vac Sci Technol B* 6:1849–1851

Rosencher E, Vinter B (2002) Optoelectronics. Cambridge University Press, Cambridge

Scherer A, Jewell J, Lee YH, Harbison J, Florez LT (1989) Fabrication of microlasers and microresonator optical switches. Appl Phys Lett 55:2724–2726

Sze S (1981) Physics of semiconductor devices, 2nd edn. John Wiley & Sons, Inc., New York

Tewordt M, Law V, Kelly M, Newbury R, Pepper M, Peacock C (1990) Direct experimental determination of the tunneling time and transmission probability of electrons through a resonant tunneling system. J Phys Condens Matter 2:896–899

Vasko FT, Kuznetsov AV (1999) Electronic states and optical transitions in semiconductor heterostructures. Springer, New York

Weisbuch C, Vinter B (1991) Quantum semiconductor structures. Academic Press, New York

Quantum Transport

16

16.1 Quantum Transport

16.1.1 The Concept of Current in Quantum Mechanics

We have seen in Sect. 16.2.1 how we could define current in classical Drude theory in terms of electrons or charges obeying Newton's law with frictional forces giving rise to resistance. In Chap. 4 we had introduced the methodology of quantum mechanics and argued that classical physics was not really the right way of looking at dynamics on a microscopic scale. In practice it turns out that the classical theory of transport is very useful indeed, and one can go a long way in understanding transport phenomena in solid-state physics and engineering using the classical method. But there comes a point beyond which the classical description does not work well anymore, and we have to consider the quantum mechanical aspects. This happens on many occasion most of which we cannot discuss here, but we can consider a very simple and common situation where quantum mechanics is needed. Consider a beam of electrons injected, for example, in the conduction band of a semiconductor via an electrode and traveling to the other electrode. Now we can ask what is the current? In classical physics, the answer is obvious if we know the velocity of the carriers. Now we can insert a potential barrier on the way, for example, a material with a higher bandgap as in Fig. 16.1, and ask: what is the resistance produced by the potential barrier on the electrons impinging on it? A classical Drude approach would obviously give us a totally oversimplified and misleading answer to this question. It would require the definition of "frictional force" but which acts only in the form of one obstacle and would not give a satisfactory picture of this well-defined and concrete transport problem. So the right starting point in this case is the *quantum mechanical definition of the current (quantum current)*. To do this, and for simplicity, we consider a one-dimensional situation and write down the continuity equation:

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_16

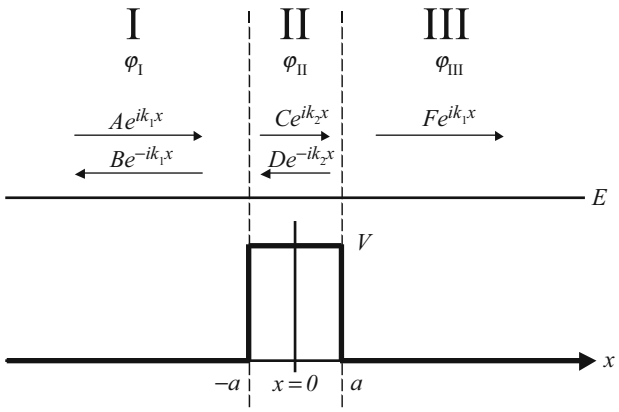


Fig. 16.1 Illustration of the three regions of particle motion

$$\frac{\partial \rho}{\partial t} + \frac{\partial J_x}{\partial x} = 0 \tag{16.1}$$

This equation is generally valid and is an expression of particle conservation where ρ is the density and J_x the current in x-direction. Now we rewrite the equation using the quantum mechanical definition of density and the time-dependent Schrödinger equation from Eq. (4.2). Recall that density in quantum mechanics is $\rho = \Psi^*\Psi$:

$$\frac{\partial(\Psi^*\Psi)}{\partial t} = \Psi^*\frac{\partial \Psi}{\partial t} + \Psi\frac{\partial \Psi^*}{\partial t} = \Psi^*\left(\frac{-i\,\widehat{H}}{\hbar}\Psi\right) + \Psi\left(\frac{i\,\widehat{H}}{\hbar}\right)\Psi^* \tag{16.2}$$

With the Hamiltonian:

$$\widehat{H} = \frac{p^2}{2m} + V(x) \tag{16.3}$$

Substituting the Hamiltonian from Eq. (16.3) into Eq. (16.2), we note that the potential energy term cancels and we have in one dimension:

$$\frac{\partial(\Psi^*\Psi)}{\partial t} = \frac{i\hbar}{2m}\left(\Psi^*\left(\nabla^2\Psi\right) - \Psi\left(\nabla^2\Psi^*\right)\right) \tag{16.4}$$

$$\frac{\partial(\Psi^*\Psi)}{\partial t} + \frac{\partial}{\partial x}\frac{\hbar}{2mi}\left(\Psi^*\left(\frac{\partial}{\partial x}\Psi\right) - \Psi\left(\frac{\partial}{\partial x}\Psi^*\right)\right) = 0 \tag{16.5}$$

From which it follows with Eq. (16.1):

$$J_x = \frac{\hbar}{2mi}\left(\Psi^* \frac{\partial \Psi}{\partial x} - \Psi \frac{\partial \Psi^*}{\partial x}\right) \tag{16.6}$$

This, Eq. (16.2), is the quantum mechanical definition of the current. It has some very interesting features. We note that it immediately follows that to carry a current, a wavefunction must be complex. In a closed system like a box, and in the absence of a magnetic field, i.e., when we have time reversal invariance, the wavefunction can always be chosen as real, and therefore the current is zero. This statement may not seem surprising perhaps, but it is very important. A plane wave e^{ikx}, for example, carries an electron current density of $-q\hbar k_x/m$ in the positive x-direction. So a beam of particles impinging from the left to right can be represented by such plane waves. Now consider what happens when a barrier is inserted in the path of such a beam. Clearly some will be reflected back, and some will go through the barrier. The question is how many per second will make it through? Classically only charges with sufficient energy to cross the barrier can go through. The quantum mechanical picture is quite different. This is one situation where one can see that the classical method does not work at all. So let us consider the quantum mechanical solution.

16.1.2 Transmission and Reflection Coefficients

Consider the diagram in Fig. 16.1 showing what happens to a beam of carriers impinging on a potential barrier.

We let a beam of particles come in from the left with amplitude A, and because some carriers will be reflected back again, there is a reflected beam with amplitude R traveling in the opposite direction. To the right of the barrier, there are no particles coming from the right, so there is a transmitted beam with amplitude F (Fig. 16.1). The potential regions are divided as follows:

$$\text{Region I} : x < -a, V = 0 \tag{16.7}$$

$$\text{Region II} : -a \leq x \leq +a, V > 0 \tag{16.8}$$

$$\text{Region III} : a < x, V = 0 \tag{16.9}$$

We also assume that the processes take place without the particles changing their energy. This is an example of "elastic scattering." So the solution of the time-dependent Schrödinger equation as defined in Chap. 4, Eq. (4.6), in each of the three regions can be written as:

$$\varphi_{\text{I}} = Ae^{ik_1 x} + Be^{-ik_1 x} \rightarrow E = \frac{\hbar^2 k_1^2}{2m} \tag{16.10}$$

$$\varphi_{\mathrm{II}} = C\mathrm{e}^{ik_2x} + D\mathrm{e}^{-ik_2x} \rightarrow E - V = \frac{\hbar^2 k_2^2}{2m} \tag{16.11}$$

$$\varphi_{\mathrm{III}} = Fe^{ik_1x} \rightarrow E = \frac{\hbar^2 k_1^2}{2m} \tag{16.12}$$

The transmission and reflection coefficients are defined by the relations:

$$T = \left|\frac{F}{A}\right|^2 \tag{16.13}$$

$$R = \left|\frac{B}{A}\right|^2 \tag{16.14}$$

To solve the problem, we now use the boundary conditions, continuity of the wavefunction, and its derivative, at $x = a$ and $x = -a$, to determine the coefficients. This gives four equations:

$$\mathrm{e}^{-ik_1a} + \frac{B}{A}\mathrm{e}^{ika_1} = \frac{C}{A}\mathrm{e}^{-ik_2a} + \frac{D}{A}\mathrm{e}^{ik_2a} \tag{16.15}$$

$$k_1\left[\mathrm{e}^{-ik_1a} - \frac{B}{A}\mathrm{e}^{ika_1}\right] = k_2\left[\frac{C}{A}\mathrm{e}^{-ik_2a} - \frac{D}{A}\mathrm{e}^{ik_2a}\right] \tag{16.16}$$

$$\frac{C}{A}\mathrm{e}^{ik_2a} + \frac{D}{A}\mathrm{e}^{-ik_2a} = \frac{F}{A}\mathrm{e}^{ik_1a} \tag{16.17}$$

$$k_2\left[\frac{C}{A}\mathrm{e}^{ik_2a} - \frac{D}{A}\mathrm{e}^{-ik_2a}\right] = k_1\frac{F}{A}\mathrm{e}^{ik_1a} \tag{16.18}$$

Solving these equations allows us to write:

$$\frac{F}{A} = \mathrm{e}^{-2ika_1}\left[\cos(2k_2a) - \frac{i}{2}\left(\frac{k_1^2 + k_2^2}{k_1k_2}\right)\sin(2k_2a)\right]^{-1} \tag{16.19}$$

$$2\frac{B}{A} = i\left(\frac{F}{A}\right)\frac{k_2^2 - k_1^2}{k_1k_2}\sin(2k_2a) \tag{16.20}$$

Using the relation:

$$\left(\left|\frac{F}{A}\right|^2 + \left|\frac{B}{A}\right|^2\right) = T + R = 1 \tag{16.21}$$

Which expresses the conservation of probability, we can find the transmission coefficient:

$$T = \frac{1}{1 + \frac{1}{4}\left(\frac{k_1^2 - k_2^2}{k_1 k_2}\right) \sin^2(2k_2 a)} \tag{16.22}$$

Or in terms of the energy E of the particle we have in the region $E > V$:

$$T = \frac{1}{1 + \frac{1}{4}\frac{V^2}{E(E-V)} \sin^2(2k_2 a)} \rightarrow E > V \tag{16.23}$$

where $k_2 = \sqrt{2m(E - V)/\hbar^2}$. The solution is still valid when $E < V$. Since k_2 is now complex, it is convenient to redefine:

$$ik_2 = \kappa \rightarrow \frac{\hbar^2 \kappa^2}{2m} = V - E > 0 \tag{16.24}$$

We can rewrite the transmission coefficient in this regime as:

$$T = \frac{1}{1 + \frac{1}{4}\frac{V^2}{E(V-E)} \sinh^2(2\kappa a)} \rightarrow E < V \tag{16.25}$$

The limit $E = V$ is of some interest. Taking this limit in Eq. (16.25) gives us:

$$T = \frac{1}{1 + \frac{2m(a)^2 V}{\hbar^2}} \rightarrow E = V \tag{16.26}$$

which shows that even when the kinetic energy is exactly as large as the potential energy, the transmission coefficient $T < 1$. Now consider a really interesting situation, namely, when $E > V$ and when in Eq. (16.23), we have the condition:

$$\sin^2(2k_2 a) = 0 \tag{16.27}$$

This happens when:

$$2ak_2 = n\pi \rightarrow n = 1, 2, 3 \ldots \tag{16.28}$$

With this condition it is equivalent to saying that when $2a = n\left(\frac{\lambda}{2}\right)2$, the transmission $T = 1$, i.e., we have a perfect transmission despite the fact that the particle has to cross an obstacle, and space is no longer homogeneous. At these resonance the particle does not see the scattering object. It behaves as if it were not there at all. The same phenomenon happens in optics for light transmission through a Fabry-Perot mirror at resonance. The requirement for perfect transmission can be rewritten as:

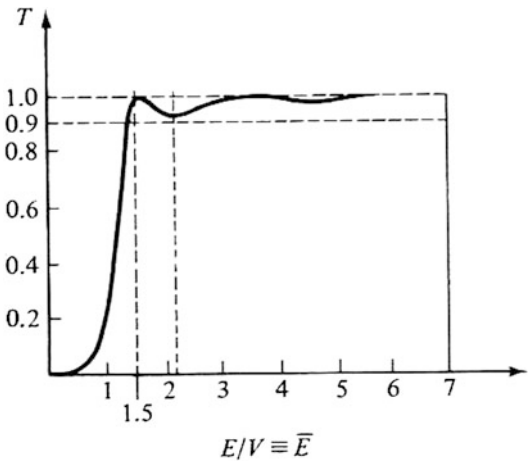


Fig. 16.2 Transmission coefficient for the single-barrier problem as a function of dimensionless energy (Reprinted with permission of Addison Wesley, R. Liboff "Quantum Mechanics, 2nd Edition" p. 231 Fig. 7.26, copyright Addison Wesley, 1992)

$$E - V = n^2 \left(\frac{\pi^2 \hbar^2}{8a^2 m} \right) \tag{16.29}$$

The behavior is shown in Fig. 16.2. When we have more than one barrier, or well, the above procedure is extended to allow reflected and transmitted waves in every region in the regions between the obstacles in a very natural generalization of the above example. In the two barrier case we have four unknown coefficients instead of two, but we have an extra boundary with two conditions.

16.1.3 Discussion

When considering the Quantum mechanical problem of transmission of particles through potential barriers, one can see a rich variety of behavior. This is true even for the simple problem of the transmission through a constant barrier as shown above. One of the most important results is the fact that one can transmit through a barrier even if one does not have enough kinetic energy to surmount it classically. This is a consequence of Heisenberg's uncertainty relation in which the delimitation of an obstacle in a specific region of space introduces indeterminance of momentum and energy. As the beam of particles is used to "measure" the presence of the object in a specific location, it no longer has a well-defined energy when it crosses the obstacle.

The transmission coefficient is, as one would expect, also a measure of the resistance or the conductance of the system. Let us examine what this implies for conductivity.

Fig. 16.3 Illustration showing the assumption that the charge reservoir on the right is shifted down by eV and that the electric field is not affecting the band structure of the system

16.1.4 The Electrical Resistance Due to Potential Barriers in Quantum Mechanics

We can consider this particle beam as being emitted from an electron reservoir in a metal and collected in a similar reservoir at a lower Fermi energy or chemical potential. So now we have the total current emitted from left to right which is (A = area, $2\ W$ = bandwidth, g_V = density of states per volume):

$$I_R = Ae \int_{-W}^{W} f(\varepsilon_k) g_V(\varepsilon_k) T(\varepsilon_k) \left(\frac{\hbar k_x}{m}\right) d\varepsilon_k \tag{16.30}$$

But from right to left, with carriers emitted from a reservoir held at a lower chemical potential of magnitude eV (see the drawing of Fig. 16.3), the current is:

$$I_L = Aq \int_{-W}^{W} f(\varepsilon_k + qV) g_V(\varepsilon_k) T(\varepsilon_k) \left(\frac{\hbar k_x}{m}\right) d\varepsilon_k \tag{16.31}$$

The net current is therefore the difference which is:

$$I_R = Aq \int_{-W}^{W} \{f(\varepsilon_k) - f(\varepsilon_k + qV)\} g_V(\varepsilon_k) T(\varepsilon_k) \left(\frac{\hbar k_x}{m}\right) d\varepsilon_k \tag{16.32}$$

For small voltages we can expand the Fermi function to order qV to obtain:

$$I = Aq^2 V \int_{-W}^{W} \left(-\frac{\partial f}{\partial \varepsilon_k}\right) g_V(\varepsilon_k) T(\varepsilon_k) \left(\frac{\hbar k_x}{m}\right) d\varepsilon_k \tag{16.33}$$

This elegant result shows how the transmission coefficient determines and defines the conductance G of the system which is:

$$G = Aq^2 \int_{-W}^{W} \left(-\frac{\partial f}{\partial \varepsilon_k}\right) g_V(\varepsilon_k) T(\varepsilon_k) \left(\frac{\hbar k_x}{m}\right) d\varepsilon_k \tag{16.34}$$

And we should remember that at $T = 0$, the derivative of the Fermi function is a delta function at the Fermi level so that $G = Aq^2 g_V(\varepsilon_F) T(\varepsilon_F) v(\varepsilon_F)$ where $v(\varepsilon_F)$ is the Fermi level velocity $v_F = \sqrt{2E_F/m}$.

This result is now easily generalized to multiples of barriers of various shapes and sizes. In all cases, we need to calculate the new transmission coefficient using the generalization of the same method.

The above derivation assumed free electrons. For band electrons the x-velocity should be replaced by $\frac{1}{\hbar}\frac{\partial \varepsilon_k}{\partial k_x}$, and the mass is the effective mass m^*.

The reader will note that if Eq. (16.34) is compared to the classical Drude result, then it is possible to define an effective Drude relaxation time via the relation $\{\sigma_{\text{Drude}} = \frac{nq^2\tau}{m^*},\ g_V(\varepsilon_F)\varepsilon_F = n\}$:

$$\tau = \frac{2TL}{v(\varepsilon_F)} \tag{16.35}$$

Note that the Drude relaxation time scales as length times $T(E_f)$. Indeed in a macroscopic resistor $T(L)$ will depend on the number of obstacles and also change, decreasing with L. If we assume that the wavefunction loses its coherence each time after scattering from an obstacle, then every time it crosses an obstacle, it is like starting again at the next one, then the resistance caused by each obstacle is additive, and the scaling of T will be as $1/L$, $1/T = \sum_n 1/T_n = \frac{L}{w} < \frac{1}{T_n} >$ where w is the average distance between the obstacles so that:

$$\tau = \frac{2\left\{< T_n^{-1} >\right\}^{-1} w}{v(\varepsilon_F)} \tag{16.36}$$

Where $<>$ = average denotes the average over the distribution of T_n. On the other hand if the system is ordered, and the total transmission coefficient does not change with L, and stays at $T = 1$, we can see that the Drude relaxation time goes to infinity with L, so that the resistivity of the macroscopic material tends to zero.

16.1.5 The Influence of the Applied Electric Field

We have up to now not mentioned the electric field. By drawing the transport path as in Fig. 16.4, we have avoided the problem of having to mention the applied electric field altogether. The only reason why there is a current is because the electrons in the right reservoir have a lower Fermi level, so the number crossing from left to right for

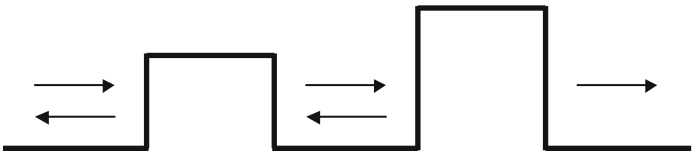


Fig. 16.4 The diagram illustrates a two barrier path with different barrier heights. The methodology of the solution is as before

a given temperature is smaller. But in reality there is, of course, an electric field gradient and electrons emitted to the right are subject to a field, and classically, they are accelerated by this field and then they scatter and/or just reach the other electrode. The quantum mechanical problem of a charge moving in an electric field was treated in Chap. 10. The problematic is not as simple as in classical physics because we deal with energy eigenstates not with acceleration and instantaneous velocities. So we avoided this issue at this stage, and it is all right to do so provided the applied potential is small, and we can work in the linear response or ohmic regime keeping only terms in first order in the applied potential. From Eq. (16.33) we observe that to linear order in V, the transmission coefficient and wavefunctions use to do derive it can be assumed to be the zero field values. So here we have neglected the fact that the field will change the energy levels, that these energy levels will have a spatial structure as treated in Sect. 16.2.1, that consequently the electrons can also relax their energy to the lattice as they move, and that there is energy dissipation or joule heating in going from left to right. The energy relaxation steps do also lead to resistance processes, but in many situations of interest in quantum devices, the barrier reflections and tunneling processes where energy is conserved are by far more important for resistance than the energy exchange with the lattice.

16.1.6 Resonant Tunneling Over a Double Barrier

Let us now consider the double barrier obstacle as shown in Fig. 16.5. In the zero-bias limit, the diagram shows the position of the Fermi level of the reservoirs by the dashed line and the two quantum well eigenstate E_0 and E_1. The application of a bias field changes the potential profile. As shown, as soon as the bias is big enough for the E_0 level to line up with the injecting Fermi level, we have the phenomenon of resonance. The incoming energy exactly matches the quantum well energy, and we have an enhanced transmission. The modeling of the transmission coefficient using the method outlined above is shown in Fig. 16.1 for two different values of well widths. One can clearly see how the current rises with bias reaches a maximum at resonance and when the injecting level and quantum well level move out alignment, the current decreases again and we have the phenomenon known as negative differential resistance (NDR). The phenomenon of transmission resonance can be understood very easily using the perturbation method of Chap. 4.

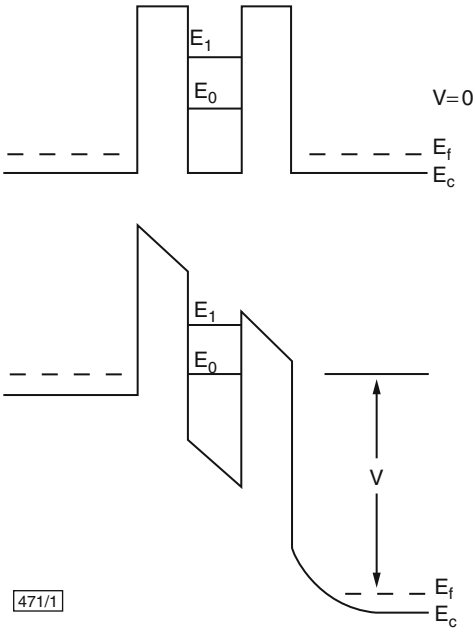


Fig. 16.5 The conduction-band profile for a double barrier resonant-tunneling structure (Copyright 1989 from "The MOCVD Challenge, Vol. 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications," Razeghi, M., p. 114, Fig. 3.37. Reproduced with permission of Routledge/Taylor & Francis Group,LLC)

We consider the electron at the injector Fermi level uncoupled to the quantum well to be in the state $\phi_I(\varepsilon_F)$ and regions I and II as defined above in Eq. (16.7) to denote the left and right reservoirs; then we allow a coupling to the quantum well level state E_0, with strength t_{I0} similarly for E_1. Now by perturbation theory to first order in the coupling, the reservoir eigenstate will be admixed to the quantum well eigenstates to give (E_0, $E_1 > \varepsilon_F$):

$$\Psi = \phi_1(\varepsilon_F) + \frac{t_{I0}}{\varepsilon_F[I] - E_0}\phi_0 + \frac{t_{I1}}{\varepsilon_F[I] - E_1}\phi_1 \tag{16.37}$$

In the present of a bias, one can approximately assume that the couplings t and the quantum well levels change so that $E_{0,\,1} \to E_{0,\,1} - qFa$ where a is the first barrier width and F the applied field so that:

$$\Psi_I = \phi_1^0(\varepsilon_F) + \frac{t_{I0}(F)}{\varepsilon_F - E_0 + qFa}\phi_0 + \frac{t_{I1}(F)}{\varepsilon_F - E_1 + qFa}\phi_1 \tag{16.38}$$

Now we see that when the levels aligns, the admixture diverges, the wavefunction acquires a high probability of mixing with the quantum well state. Of course at resonance one has to use the degenerate state perturbation method also explained in Chap. 4. To evaluate the full transmission with this method, one has to allow the coupling to the second reservoir as well and then obtain the amplitude of the initial state in the final state. But this is straightforward. Experimentally one can observe these negative differential resistance resonances, but it is not a trivial task, and a

number of conditions have to be satisfied. The quantum well levels are in practice not just sharp eigenstates but they are subject to broadening in the plane specially if the semiconductor layers are highly doped. Note that the same perturbation procedure can be applied to the second reservoir, and thus one can couple the initial reservoir to the final reservoir by a simple extension of the above perturbation theory, i.e., replacing ϕ_0 with:

$$\phi_0 = \phi_0^0 + \sum_{\varepsilon_{\mathrm{II}}} \frac{t_{0\varepsilon}^{w}(F)}{E_0 - \varepsilon_{\mathrm{II}} + qFa} \phi_{\mathrm{II}}(\varepsilon) \tag{16.39}$$

where $\varepsilon_{\mathrm{II}}$ is now any energy in the final reservoir and $t_{0\varepsilon_{\mathrm{II}}}^{w}$ is the admixture energy from the zero energy state in the well w to the final energy in the right reservoir, or electrode II, and not just the Fermi level. When the carrier has arrived in the second reservoir, it still has the field energy which it has acquired on the way, and it can deposit it in the electrode. Similarly for the upper energy level in the central well. The scattering-induced broadening washes out some of the features as one might expect. For example, from Eq. (16.38), the broadening can be represented as an imaginary part contribution $i\Gamma$ to the energy which gives the admixture probability:

$$|a_{I0}|^2 = \frac{|t_{I0}|^2}{(\varepsilon_F - E_0 + qFa)^2 + \Gamma^2} \tag{16.40}$$

Then there is also the thermal broadening effects which we can also include in the broadening width, and the fact that off resonance, when the quantum inelastic tunneling transport paths are improbable, the carrier can cross the obstacle by using the thermal activation into the quantum well levels. The so-called phonon-assisted pathways are important as we go up in temperature. Figure 16.6 shows the

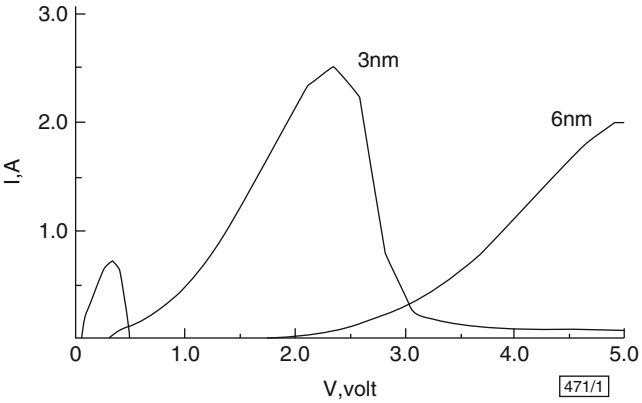
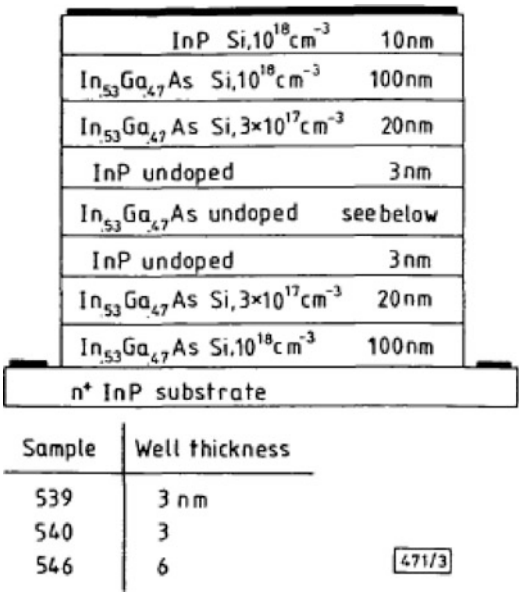


Fig. 16.6 Resonant-tunneling current/voltage simulations for 3- and 6-nm-wide wells (Copyright 1989 from "The MOCVD Challenge, Vol. 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications," Razeghi, M., p. 114, Fig. 3.38. Reproduced with permission of Routledge/Taylor & Francis Group, LLC)



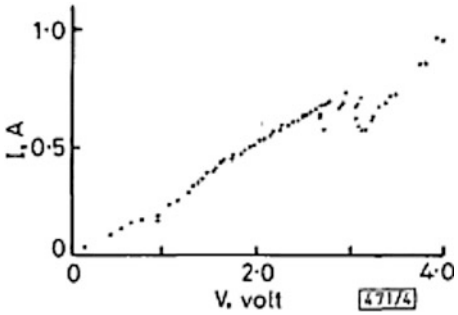


Fig. 16.7 The structure of the device used by Razeghi et al. in 1987 (Reprinted with permission from Electronics Letters Vol. 23, Razeghi, M., Tardella,A., Davis, R., Long, A., and Kelly, M., "Negative Differential Resistance at room temperature from resonant-tunneling GaInAS/InP double barrier structures," p. 116. Copyright 1987, IEEE)

Fig. 16.8 Pulsed current/voltage characteristic for a sample showing negative differential resistance at 3 V bias (Reprinted with permission from Electronics Letters Vol. 23, Razeghi, M., Tardella, A., Davi, R., Long, A., and Kelly, M., "Negative Differential Resistance at room temperature from resonant-tunneling GaInAS/InP double barrier structures," p. 116. Copyright 1987, IEEE)

calculated negative differential resistance in the device structure shown in Fig. 16.7. In Fig. 16.8 one can see the negative differential resistance obtained by a voltage pulsed technique at room temperature. The pulsing ensures that space charge and thus internal field effects do not mask the quantum tunneling process.

It is also possible to see negative differential resistance (NDR) in the steady state in some materials and especially if we go down to low temperatures where thermal

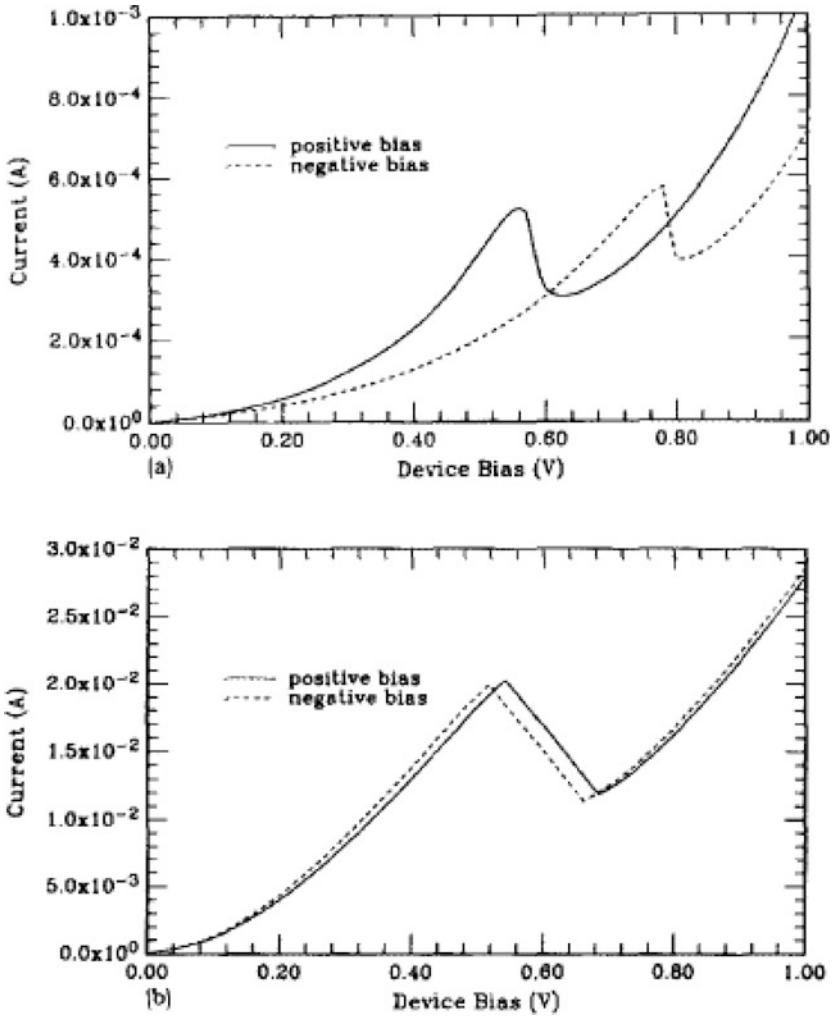


Fig. 16.9 Resonant tunneling through a double GaAs/AlAs superlattice barrier, single-quantum well heterostructure (Reprinted with permission from Applied Physics Letters Vol. 49, Reed, M., Lee, J., Tsai, H., "Resonant tunneling through a double GaAs /AlAs superlattice barrier single quantum well heterostructure," p. 158. Copyright 1986, American Institute of Physics)

broadening and pathways are suppressed. Figure 16.9 shows the NDR produced at room temperature in a GaAs/AlAs double barrier system.

The reader is referred to the books by M. Kelley and C. Weisbuch and B. Vinter for a more detailed review of negative differential resistance in quantum devices.

16.1.7 The Superlattice Dispersion

One of the most interesting phenomena of quantum physics is produced when we apply a strong electric field to narrow energy bands. The case of free electrons in an electric field was treated in Chap. 11, and there we were interested in what happens to the optical absorption in a semiconductor. Here we want to focus on what happens in narrow energy bands and show that the physics is very novel and interesting. Let us first recall the Kronig-Penney band structure of Chap. 5. This model is actually a very good representation of the band structure of a semiconductor superlattice as shown schematically in Fig. 16.10. A popular example considered in practically every specialized book (see Razeghi 1989; Weisbuch and Vinter 1991, Davis 2000, Kelly 2000) is the GaAs/AlAs superlattice.

With infinite barriers, the individual quantum wells would have confined eigenstates as shown in Fig. 4.7 in Chap. 4. When the barriers are finite, the quantum well confined levels can tunnel across and then overlap with each other thus forming energy band whose width can be adjusted as was done in the Kronig-Penney model in Sect. 16.2.1. In principle we can proceed as for the Kronig-Penney model here too. But there is a quicker way to examine the band structure specially when there is an electric field and that is to use the "tight binding model." In the framework of the tight-binding model discussed in Chaps. 1 and 5, the overlap or coupling between the confined levels of subband "n" in two adjacent wells can be written as:

$$\begin{aligned} t^n_{l,l+1} &= \int dz \Psi^*_{n,l} V(z)) \Psi_{n,l+1} \\ t_{l,l+1} &= t^*_{l+1,l} \end{aligned} \tag{16.41}$$

where Ψ is the confined quantum well state in well l, $V(z)$ is the potential caused by the adjacent well which starts at $z = a$ and finishes at $z = b$, and the superlattice distance is $c = \frac{a+b}{2}$ (see Fig. 16.10). The coupled Bloch wavefunction of energy E in one dimension can be written as a superposition of the individual quantum well (n subband) state:

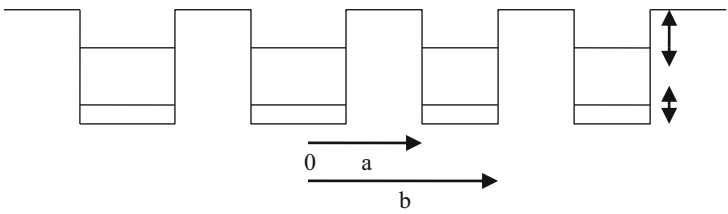


Fig. 16.10 The schematic representation of a superlattice periodic potential in the growth direction. The vertical arrows denote the expected width of the so-called superlattice minibands formed by the coupling of the quantum well wavefunctions as in the Kronig-Penney model of Sect. 16.2.1

$$\phi_E(z) = \sum_l C_l^n \Psi_{n,l} = \sum_l C_l^n |n,l\rangle \tag{16.42}$$

Taking the expectation value of the Hamiltonian with the coupled states, Eq. (16.42) gives by definition:

$$E = \langle \phi_E | H_0 + V | \phi_E \rangle \tag{16.43}$$

When substituting Eq. (16.42) into Eq. (16.43) and evaluating the right-hand side, we encounter terms of the form:

$$\langle n,l|H_0|n,l\rangle = E_n^0 \tag{16.44}$$

$$\langle n,l|H_0|n,l'\rangle \tilde{} 0 \rightarrow l \neq l' \tag{16.45}$$

$$\langle n,l|H_0|n',l\rangle = 0 \rightarrow n \neq n' \tag{16.46}$$

$$\langle n,l|V(z)|n,l\pm 1\rangle = t \tag{16.47}$$

$$\langle n,l|V(z)|l',n'\rangle \tilde{} 0 \rightarrow l' \neq l \pm 1, n \neq n' \tag{16.48}$$

The above relations give the following relations for the coefficients:

$$\begin{aligned} EC_l^n &= E_n^0 C_l^n + t_n\left(C_{l+1}^n + C_{l-1}^n\right) \\ t_{l,l\pm 1}^n &= t_n \end{aligned} \tag{16.49}$$

The equation can be solved by making use of the translational symmetry and Bloch's theorem which says that:

$$\begin{aligned} C_{l+1}^n &= C_l^n \mathrm{e}^{ik_z c} \\ C_{l-1}^n &= C_l^n \mathrm{e}^{-ik_z c} \end{aligned} \tag{16.50}$$

where c is the lattice repeat distance so Eq. (16.49) becomes:

$$\begin{aligned} \left(E - E_n^0 - 2t_n \cos k_z c\right) C_l^n &= 0 \\ E_n(k_z) &= E_n^0 + 2t_n \cos k_z c \end{aligned} \tag{16.51}$$

This cosine dispersion is a good approximation to the superlattice band structure in the growth direction. The wavefunctions in Eq. (16.49) are now labeled with the k_z vector as $\varphi_{n,kz}$.

16.1.8 The Stark-Wannier States

Now consider the application of an electric field in the z-direction. This introduces an extra term in the Hamiltonian, with F denoting the applied electric field:

$$H_F = H + qFz \tag{16.52}$$

Now we can expand the new eigenstates in terms of the new Bloch states we have just derived. For convenience, we only deal with z-component, the x,y directions are free electron effective mass states, and the total energy is decomposed as $E = E_z + \varepsilon(k_y) + \varepsilon(k_x)$

$$\Psi(z) = \sum_{n,k_z} A_n(k_z)\phi_{n,k_z}(z) \tag{16.53}$$

With $H_F\Psi(z) = E_z\Psi(z)$, we have after multiplying the left with using the orthogonality of the wavefunctions:

$$\left(E_z - E_{n,k_z}\right)A_n(k_z) = qF\sum_{k_z',j} Z_{n,k_z;j,k_z'}A_j\left(k_z'\right) \tag{16.54}$$

where:

$$Z_{n,k_z;j,k_z'} = \int dz\phi^*_{n,k_z}(z)z\phi_{j,k_z'}(z) \tag{16.55}$$

and the matrix elements Eq. (16.55) obey the rule:

$$Z_{n,k_z;j,k_z'} = i\delta_{n,j}\delta_{k_z,k_z'}\frac{\partial}{\partial k_z} \tag{16.56}$$

Substituting Eq. (16.56) back into Eq. (16.54) gives us a first-order differential equation:

$$\left(E_z - E_{n,k_z}\right)A_n(k_z) = iqF\frac{\partial A_n(k_z)}{\partial k_z} \tag{16.57}$$

which we can integrate straight away by first dividing throughout with A_n, to give after substitution Eq. (16.53):

$$\Psi_\nu{}^n(z) = \int_{-\pi/c}^{\pi/c} c^{1/2}\frac{dk_z}{2\pi}\exp\left[ik_z(z-\nu c) - \frac{it_n}{qFc}\sin k_z c\right] \tag{16.58}$$

where we have used the relations:

$$E_z = qFc\nu + E_n^0 \tag{16.59}$$

$$\sum_{k_z} \rightarrow \frac{L}{2\pi}\int dk_z \tag{16.60}$$

$$E_n(k_z) = E_n^0 + 2t_n \cos k_z c \tag{16.61}$$

This defines the energy levels. The Bloch symmetry in the superlattice which stipulates that the same functions must be reproduced if the origin is shifted to an equivalent site forces the indices ν to be integers ranging from $(-\infty, \infty)$. The complete wavefunctions and energies are now given by:

$$\Psi_E(x,y,z) = \int_{-\pi/c}^{\pi/c} c^{1/2}\frac{dk_z}{2\pi}\exp\left[ik_z(z-\nu c) - \frac{2it_n}{qFc}\sin k_z c\right] \times \left(\frac{1}{L_x L_y}\right)^{1/2}\exp\left(ik_x x + ik_y y\right) \tag{16.62}$$

$$E = qFc\nu + E_n^0 + \frac{\hbar^2}{2m^*}\left(k_x^2 + k_y^2\right) \rightarrow -\infty < \nu < \infty \tag{16.63}$$

The wavefunction in the z-direction is a well-known function in mathematical physics, and it is called the Bessel function of the first kind J and can be written in the usual way as:

$$\int_{-\pi/c}^{\pi/c} c\frac{dk_z}{2\pi}\exp\left[ik_z(z-\nu c) - \frac{it_n}{qFc}\sin k_z c\right] = J_{\frac{z}{c}-\nu}\left(\frac{2t_n}{qFc}\right) \tag{16.64}$$

The new physics is fascinating. The first thing we note is that the Bessel functions are localized in the z-direction. Starting from an origin νc, the wavefunction decays after a distance of L_n where

$$L_n = \frac{2t_n}{qF} \tag{16.65}$$

The stronger the electric field, the smaller the wavefunction. This is classically completely counterintuitive. The eigenstates will form a so-called Stark-Wannier (SW) ladder centered about the middle of every quantum well in the superlattice, extending to about the distance L_n in the z-direction. Since the banding parameter t_n increases as we go up in the quantum well subband index n, because the confinement is weaker at higher energies (see Chap. 4 Eq. (4.47)), the localization length of the Stark-Wannier states increases for the higher energy bands in the superlattice.

If an electron is put in any of the Stark-Wannier states, it is in an eigenstate and will stay there forever and not transport charge unless it is allowed to couple with the photon or phonon fields. And of course it is coupled to those fields, and thus it will relax down the Stark-Wannier energy ladder emitting light or heat as it moves. The current or drift velocity is then limited by the rate of energy relaxation. If we calculate this rate (Movaghar 1987), then we can calculate the drift velocity. If the energy relaxing coupling Hamiltonian is $H_r(z)$ then by the Fermi golden rule (see Chap. 10), we have the rate:

$$\begin{aligned}\gamma_{\nu\nu'} &= \frac{2\pi}{\hbar}\sum_s \left|\int dz \Psi_\nu^*(z) V_r(z,\omega_s)\Psi_{\nu'}\right|^2 \delta(qcF\nu' - qcF\nu + \hbar\omega_s)(1+n(\omega_s)) \\ &\quad \rightarrow qcF\nu > qcF\nu' \\ V_r(z,t) &= V_r(z,\omega_s)\mathrm{e}^{i\omega_s t} + V_r^*(z,\omega_s)\mathrm{e}^{-i\omega_s t}\end{aligned} \tag{16.66}$$

where the excitations ω_s emitted can be phonons or photons and V_r is the relaxation coupling. Even though the dominant relaxation channel is phononic, specially when there is a resonant Stark-Wannier transition with an optic phonon mode, the photons emission can be turned into a lasing emission in the so-called quantum cascade laser (QCL) structure (see Faist et al. 1994 and Slivken et al. 2002, in the further reading section for more information). In the process $\nu \rightarrow \nu'$ the distance traveled is $d = c(\nu' - \nu)$ which then defines a transfer velocity. Moving against the field is also possible but will involve an absorption of the excitation involved. But the excitation has to be available so the "up rate" involves an activation factor via the phonon or photon occupation number $n(\omega_s)$. Again from the Fermi golden rule for absorption going up the energy ladder, we have:

$$\begin{aligned}\gamma_{\nu\nu'} &= \frac{2\pi}{\hbar}\sum_s \left|\int dz \Psi_\nu^*(z) V_r(z,\omega_s)\Psi_{\nu'}\right|^2 \delta(qcF\nu - qcF\nu' + \hbar\omega_s)(n(\omega_s)) \\ &\quad \rightarrow qcF\nu' > qcF\nu\end{aligned} \tag{16.67}$$

$$n(\omega_s) = \frac{1}{(e^{\hbar\omega_s/k_bT} - 1)} \tag{16.68}$$

We conclude that the motion of electrons in finite energy bands in quantum mechanics gives rise to a very different physical picture than in classical physics. What is happening here viewed in the semiclassical picture is that the carriers are accelerated by the electric field to go up the Bloch band energy ladder (assume $t_n < 0$):

$$\frac{d\vec{k}}{dt} = -q\vec{F} \tag{16.69}$$

$$E_n(k) = E_n^0 - 2|t_n| \cos k(t)c \tag{16.70}$$

Since the energy is periodic, in the absence of energy dissipation, the electron accelerates and then slows down again when it reaches the top of the energy band and then actually has to return in space to move against the field. It cannot go further than a certain distance in space without losing energy, because its energy will have reached the top of the allowed energy band which is $E = E_n^0$. Its motion in energy space is periodic and so is its motion in real space. This motion is called Bloch oscillations, and one can understand that the particle is being actually localized by the applied field. In order to move a longer distance than the Stark-Wannier length, the carrier has to emit energy, and in the quantum mechanics point of view, relax down into the adjacent Stark-Wannier state or classically speaking, start a new journey forward in field direction in space as soon as it has emitted energy, thus avoiding to have to go back in space. The reason why the simple classical viewpoint is valid in many situations is that the electron has in most cases plenty of opportunity to relax its energy before it has reached the top of the band. In formal language, the energy relaxation time (inverse of the rate) is short compared to the time it takes to reach the top of the band via Eqs. (16.69) and (16.70). Under these circumstances it is possible to think of the relaxation as a frictional force acting on a carrier which remains in the effective mass regime staying in the small k region with $\cos kc \sim 1 - (kc)^2/2$.

16.1.9 Quantum Transport in Two-Dimensional Channels

One of the most brilliant discoveries of quantum well physics is the idea of the modulation-doped structure, shown on the right-hand side of Fig. 16.11 and more explicitly in Fig. 16.13. The dopant is introduced in the barrier layer as, for example, in Fig. 16.11 in the AlGaAs layer. In this way the carrier moves into the region of low energy which is the conduction band of the GaAs layer and leaves its counterion behind in the barrier. The counterion is now well separated in space from the conducting channel formed in parallel and shown in Fig. 16.13. The physical separation from the dopant means that the charge impurity scattering contribution is considerably reduced compared to the normal case.

The scattering rate can be written as:

$$\frac{1}{\tau_{\text{imp}}} = N_{\text{imp}} \frac{m^*}{2\pi\hbar^3 k_F^3} \left(\frac{q^2}{2\varepsilon_0 \varepsilon_b}\right) \int_0^{2k_F} \frac{\mathrm{e}^{-2k|d|}}{[k + q_I G(k)]^2} \left(\frac{b}{b+k}\right)^6 \frac{k^2}{\sqrt{1 - (k/2k_F)^2}} dk \tag{16.71}$$

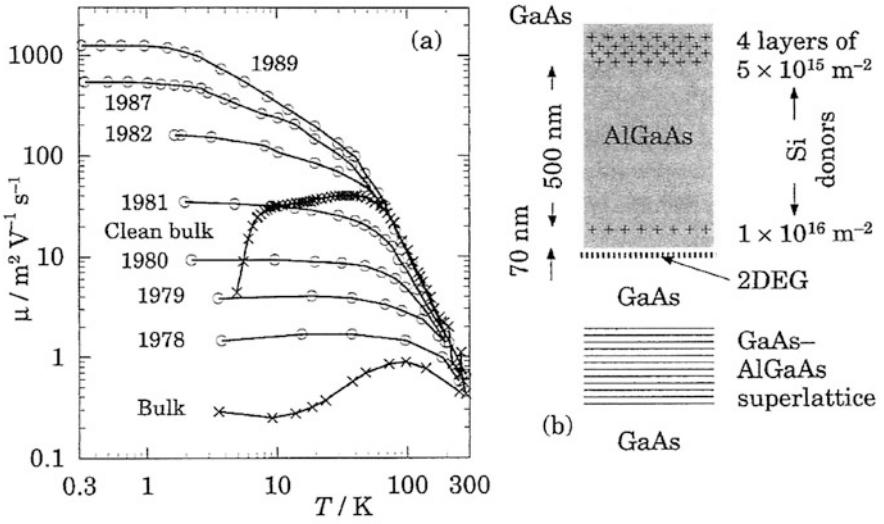


Fig. 16.11 Mobility of various two-dimensional electron gases 2DEGs as a function of temperature (circles) showing how the peak mobility which is limited by impurity scattering has increased over the 20 years shown. The mobility of bulk samples is shown for comparison (crosses) for older material (bulk) and newer material (clean bulk). On the right we see a simplified structure of a wafer grown in the sample of highest mobility (From J.H Davis "The physics of low dimensional semiconductors" p. 360, Fig. 9.11a, copyright Cambridge University press 1998, redrawn from Applied Physics Letters Vol. 55 "Electron mobilities exceeding 10^7 $cm^2/V{\cdot}s$ in modulation-doped GaAs," pg. 1888, Fig. 1, copyright American Institute of Physics 1989. Reprinted with permission of Cambridge University press and American Institute of Physics)

$$G(k) = \frac{1}{8}\left[2\left(\frac{b}{b+k}\right)^3 + 3\left(\frac{b}{b+k}\right)^2 + \left(3\frac{b}{b+k}\right)\right] \tag{16.72}$$

where b is defined below, N_{imp} and N_{2D} are the 2D-impurity and the 2D electron concentrations in the dopant layer, respectively, and d is the distance to the impurity dopant layer as measured from the edge of the GaAs layer (see Fig. 16.12); the remaining parameters are defined later (Fig. 16.13).

And the wavefunction in the channel (see Fig. 16.12) is plane wave like in (x,y) plane and in confined z-direction-given by (Ando et al. 1982):

$$u(z) = \left(\frac{b^3}{2}\right)^{1/2} z\exp[-bz/2] \tag{16.73}$$

b is the so-called Fang-Howard decay parameter given by:

$$b = \left(\frac{33m^* q^2 N_{2d}}{8\hbar^2 \varepsilon_0 \varepsilon_b}\right)^{1/3} \tag{16.74}$$

We also have the definitions for Eq. (16.71):

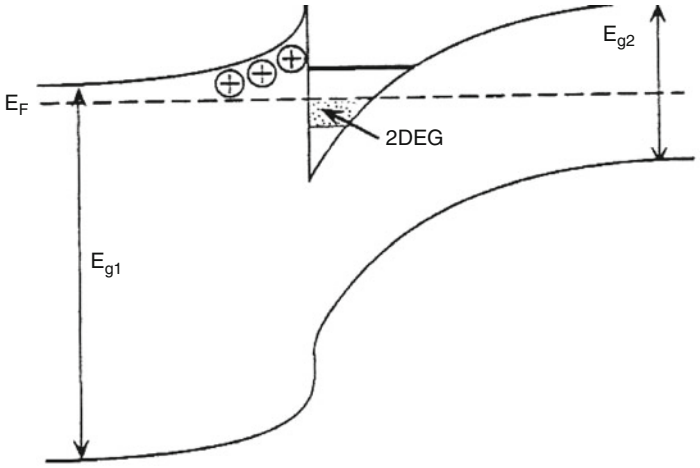


Fig. 16.12 Energy band diagram for a modulation-doped heterostructure (Copyright 1995 from "The MOCVD Challenge Vol. 2, A survey of GaInAsP-GaAs for photonic and electronic device applications," Razeghi, M., p. 371, Fig. 9.2. Reproduced with permission of Routledge/Taylor & Francis Group, LLC)

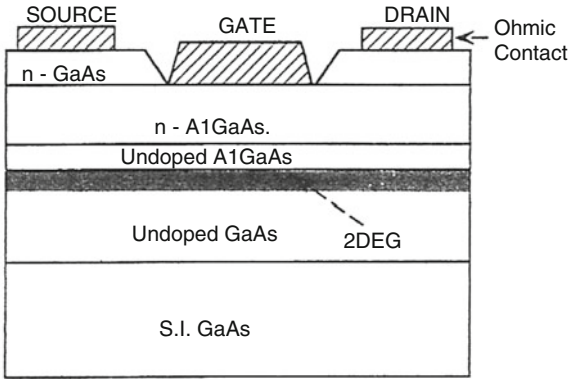


Fig. 16.13 Typical cross section of an AlGaAs/GaAs MODFET (Copyright 1995 from "The MOCVD Challenge Vol. 2, A survey of GaInAsP-GaAs for photonic and electronic device applications," Razeghi, M., p. 372, Fig. 9.3. Reproduced with permission of Routledge/Taylor & Francis Group, LLC)

$$q_I = 2/a_{\mathrm{BI}}$$
$$a_{\mathrm{BI}} = \frac{4\pi\varepsilon\varepsilon_0}{m^* q^2} \tag{16.75}$$

where a_{BI} is the effective radius around the hydrogenic dopant charge. To a reasonably good approximation, the rate can be written in the very simple form:

$$\frac{1}{\tau_{\mathrm{imp}}} = N_{\mathrm{imp}} \frac{\pi\hbar}{8m^*(|d|k_F)^3} \tag{16.76}$$

For a concentration N_{2D}~ 3.10^{15} m^{-2}, $d = 30$ nm and $N_{\text{imp}} = 10^{16}$ m^{-2}, $m = m_0$ The exact result is μ~50m^2/Vs and if this is the dominant source of scattering at low temperatures, the mean-free path is $l_c \sim 5$ μm.

In contrast, the acoustic phonon scattering rate, using the same wavefunctions, can be shown to be using the electron-phonon coupling from Chap. 15:

$$\frac{1}{\tau_{\text{ac}}} = \frac{3m^* bkT(D_{\text{ac}})^2}{16\rho_d v_s^2 \hbar^3} \tag{16.77}$$

The LO optic phonon rates start at higher temperatures when such phonons can be excited as expressed by the Bose distribution function and give:

$$\frac{1}{\tau_{\text{LO}}} = \left(\frac{2m^*\omega_{\text{LO}}}{\hbar}\right)^{1/2} \frac{q^2}{8\hbar\varepsilon_0}\left(\frac{1}{\varepsilon(\infty)} - \frac{1}{\varepsilon(0)}\right)\left(\frac{1}{\exp\left(\frac{\hbar\omega_{\text{LO}}}{k_b T}\right) - 1}\right) \tag{16.78}$$

In GaAs the LO-phonon scattering rate is ~ $\{10^{13}\frac{1}{e^{\hbar\omega_{\text{LO}}/k_b T}-1}\text{Hz}\}$ and therefore very strong and dominant when $T > 40$ K. The temperature structure is very close to what is shown in Figs. 16.11 and 16.14 as one crosses into the temperature region when optic phonons are excited, i.e., $T > 40$ K. Figure 16.12 defines the two-dimensional electron gas referred to as 2DEG, exhibits the difference between the mobility in the bulk and in a 2DEG gas, and shows how the various theoretical scattering mechanisms can explain the temperature behavior in the two systems.

16.1.10 Motion in the Plane: Magnetoresistance and Hall Effect in Two-Dimensional Electron Gas

As a consequence of the high mobilities and long mean free paths, carrier dynamics also exhibit beautiful quantum effects in the presence of an applied magnetic field. We first recall what happens to the spectrum of a 2DEG in a magnetic field from Chap. 4. As we recall from Fig. 4.14 and Eq. (4.174), the magnetic field enhances the Landau level splitting and increases the density of states in each subband, so that as we go up in magnetic field B, the Fermi level is pushed down until at very high B field, only the first subband is occupied. As the Fermi level moves through the Landau levels, one must remember that these have a finite broadening caused by scattering from disorder, and they will typically look like Fig. 16.15 or Fig. 16.16. The localization of eigenstates at the band edges produced by disorder and the Shubnikov-de Haas oscillations in the conductivity and the quantum Hall conductivity derived from the Quantum hall effect (QHE) were discussed briefly and qualitatively in Chap. 15. The objective in this chapter is to introduce the reader also to the new mathematical physics.

Consider first the classical Drude magnetoresistance we derived in Sect. 10.8. This formula was good enough for most optical applications, but it ceases to be

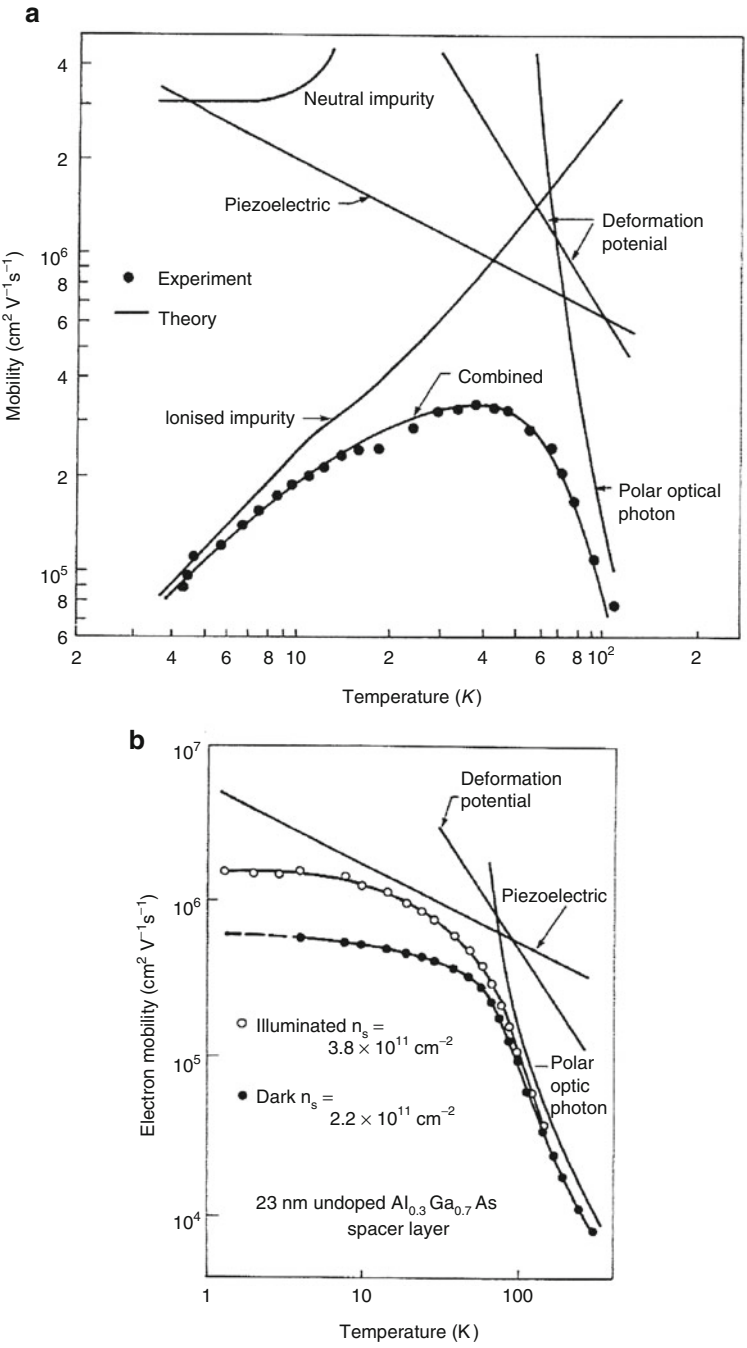


Fig. 16.14 Theoretical and experimental mobility of (**a**) high-purity GaAs (**b**) at a heterojunction. The theoretical lines give the limits placed on the mobility by the relevant scattering mechanisms (Part (**a**) Reprinted with permission from Thin Solid films Vol. 31, G E Stilman and C M Wolfe, "Electrical characterization of thin epitaxial layers" pg. 69 Fig. 7, copyright 1976 Elsevier. Part (**b**) Reprinted with permission from IEEE Tech Digest, International Electron Devices Meeting, Dilorenzo, J., Dingle, R., Feuer, R., Gossard, M., Hendel, A., Hwang, R "Materials and devices Characterization for selectively doped heterojunction transistors," p. 578–581, Fig. 3. Copyright 1982, IEEE, New York)

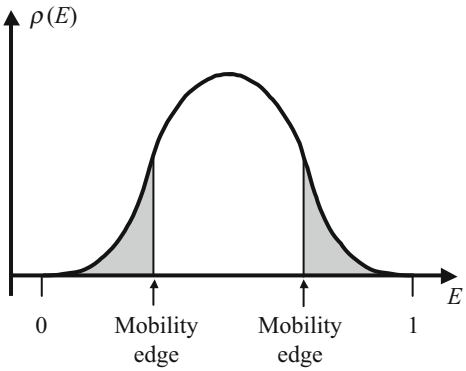


Fig. 16.15 Density of states model of a single Landau level broadened by disorder and showing the shaded region where the eigenstates are expected to be localized. As the Fermi level crosses the density of states, it goes through extended and localized regions. The longitudinal conductivity is zero when the Fermi level is in the localized region, and the Hall conductivity does not change until the Fermi level is again in extended states

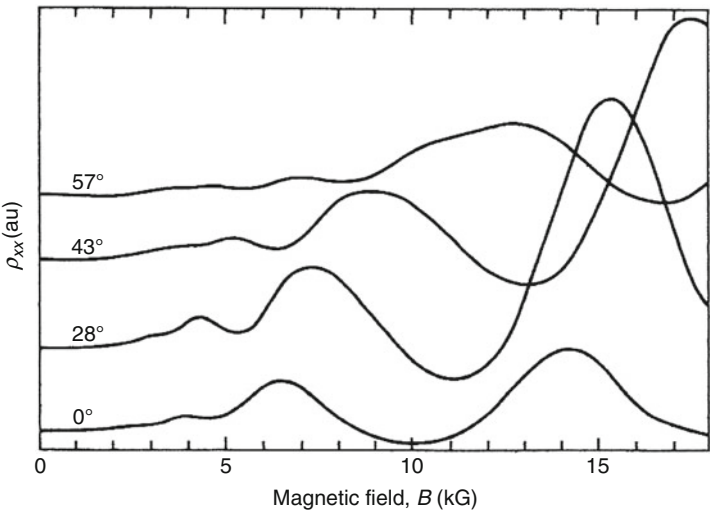


Fig. 16.16 Longitudinal magnetoresistance ρ_{xx} of low-pressure MOCVD-grown GaInAs/InP versus magnetic field for various angles at 4.2 K (Copyright 1989 from "The MOCVD Challenge, Vol. 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications," Razeghi, M., p. 125, Fig. 3.48. Reproduced with permission of Routledge/Taylor & Francis Group, LLC)

useful at high magnetic fields in high mobility 2DEG systems, where Landau levels appear. An improvement which takes the Landau quantization into account and allows also for a lifetime broadening and impurity scattering was derived by Ando et al. (1982). The conductivity becomes in the effective mass approximation:

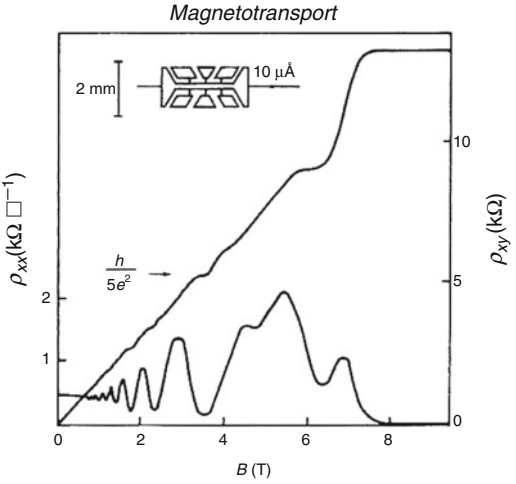


Fig. 16.17 Resistivity data ρ_{xx} and ρ_{xy} at 1.3 K for an InGaAsInP heterostructure (MOCVD) growth technique ($4.3 \times 10^{11}\text{cm}^{-2}$) $\mu = 60000\,\text{cm}^2/\text{V/s}$ (Copyright 1989 from "The MOCVD Challenge, Vol.1:A Survey of GaInAsP-InP for Photonic and Electronic Applications," Razeghi, M., p. 131, Fig. 3.53. Reproduced with permission of Routledge/Taylor & Francis Group, LLC)

$$\sigma_{xx}^{2D}(\omega = 0, B) = \frac{q^2 N_{2D}\tau}{m^*}\frac{1}{1+(\omega_c\tau)^2}\left\{1 - \frac{2(\omega_c\tau)^2}{1+(\omega_c\tau)^2}S(\omega_c)\right\} \tag{16.79}$$

$$S(\omega) = \frac{2\pi^2 k_b T}{\hbar\omega_c}\text{cosech}\left(\frac{2\pi^2 k_b T}{\hbar\omega_c}\right)\cos\left(\frac{2\pi\varepsilon_F}{\hbar\omega_c}\right)\exp\left[-\frac{\pi}{\omega_c\tau}\right] \tag{16.80}$$

when $(\omega_c\tau < 1)$, where τ is the scattering rate ω_c the cyclotron frequency and N_{2D} the carrier density in 2D. The oscillatory structure of this function reproduces the measured structure shown in Figs. 16.16 and 16.17, and indeed it is possible from this comparison to deduce the effective mass and scattering rates (Weisbuch and Vinter 1991).

Now consider the quantum Hall conductivity also shown in Fig. 16.17. In order to fully appreciate the significance of this phenomenon, it is necessary to mentally get rid of the scattering process altogether and consider the pure quantum state with magnetic field B and in the presence of the applied electric field F. The full Hamiltonian in the Landau gauge is:

$$H\Psi = \left\{\frac{1}{2m}(p_y + qxB_z)^2 + \frac{p_x^2}{2m^*} + \frac{p_z^2}{2m^*} + qF_x x\right\}\Psi = E\Psi \tag{16.81}$$

In a 2DEG formed by modulation doping or a quantum well confinement, the motion in the z-direction is bounded, so that the wavefunctions and energies in the z-direction are not free-electron like. But for the present purpose, this is immaterial. We also drop the z-index on the B field. The new electric field-dependent term qFx can be combined into the x-dependent B-terms. The solution is a straightforward extension of the $F = 0$ problem to give:

$$\Psi_n(p_y,p_z) = \left(\frac{1}{L_y L_z}\right)^{1/2} \exp\left[\frac{i}{\hbar}(p_y+p_z)\right]\phi_n\left(x-x_{p_y}\right) \tag{16.82}$$

$$\phi_n(x) = \left(\frac{1}{2^n n!\sqrt{\pi}}\right)^{1/2} \frac{1}{\sqrt{a_0}} \exp\left[-\frac{1}{2}(x/a_0)^2\right] H_n(x/a_0) \tag{16.83}$$

$$a_0 = \left(\frac{\hbar}{m^*\omega_c}\right)^{1/2} \tag{16.84}$$

The energy levels are:

$$E_n(p_y,p_z) = \hbar\omega_c(n+1/2) + qF_x x_{p_y} + p_z^2 + \frac{m^*}{2}(F_x/B)^2 \tag{16.85}$$

$$x_{p_y} = \frac{1}{qB}\left[p_y + m^*\left(\frac{F_x}{B}\right)\right] \tag{16.86}$$

2DEG confinement in the z-direction would replace $p_z^2 \rightarrow E_{z,\mu}$, and these would correspond to the z-eigenvalues of the Fang-Howard model in Eq. (16.73), for example. Note the interesting fact that the energy levels depend on the value of the y-momentum and that the value is asymmetric with p_y, so that the negative p_y will have lower energy, and these eigenstates will be the first to be occupied at $T = 0$. Now consider the quantum current in the y-direction. Remember that the velocity in a magnetic field is different so that the momentum becomes $mv_y = -i\hbar\frac{\partial}{\partial y} \rightarrow -i\hbar\frac{\partial}{\partial y} + qB_z x$, now taking the matrix element of the new operator and then giving a velocity which is remarkably independent of p_y:

$$J_y = -q\sum_{p_y} f(p_y)\frac{1}{B}F_x = -N_{2D}\frac{q}{B}F_x \tag{16.87}$$

The sum over p_y at $T = 0$ just gives the total number of occupied electronic states. Noting that the degeneracy of each Landau level is $\frac{qB}{h}$, so that for i_L number of full bands $N_{2D} = i_L qB/h$, we obtain for the Hall conductivity of electrons:

$$\sigma_{xy} = -q\sum_{p_y} f(p_y)\frac{1}{B} = -i_L\frac{q^2}{h} \tag{16.88}$$

The Hall conductivity is dependent only on the filling index of the Landau subbands i_L. The quantity one normally works with is the Hall resistivity defined as:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_{xy}^2 + \sigma_{xx}^2} \tag{16.89}$$

and the magnetoresistivity is given by:

$$\rho_{xx} = \frac{\sigma_{xx}}{\sigma_{xy}^2 + \sigma_{xx}^2} \tag{16.90}$$

When the Fermi level is in a Landau gap, $\sigma_{xx} = 0$ but σ_{xy} is not, so the magnitude of the Hall resistance is:

$$\rho_{xy} = \frac{1}{i_L}\frac{h}{q^2} \tag{16.91}$$

Now let us look at an experiment. We note that Fig. 16.17 is similar to Fig. 15.20 obtained in a more accurate measurement configuration. It shows that the measured Hall conductivity has plateaus. Every time the Fermi level passes into the edge region where the levels are localized as shown in Fig. 16.15, the Hall resistivity does not change, because localized levels do not conduct. But as soon as we pass the edge of the Landau levels, and reach the middle of the Landau bands where the levels are delocalized, the Hall conductivity changes again, going "up or down" depending on whether we are decreasing or increasing the magnetic field. The Hall resistivity decreases rapidly as we cross the "free" region in the direction of increasing magnetic field B. Now consider the longitudinal resistivity ρ_{xx}. Every time the Fermi level passes through the Landau gaps, we know from Eq. (10.22) in Chap 10 that a null or a localized spectrum implies a zero conductivity and thus from Eq. (16.91) a peak in the resistance. Since a minimum scattering rate finite relaxation time is inevitable, the Ando formula (Eqs. (16.79) and (16.80)) is in practice a good approximation. The remarkable feature of the Hall conductivity is that despite the fact that there are regions of localized states and that electrons that occupy these states do not conduct, the total Hall current behaves as if all the electrons have transported normally as free carriers. The free carriers have acquired a higher velocity which exactly compensates for the ones which are not moving (Prange 1981).

The important point is that the quantum Hall current flows even when the band is full, and the Fermi level is in a mobility gap in the bulk of the sample. It is not a Fermi level property as is ordinary conduction. Indeed when the Fermi level is in a gap, by definition, all the $\{+p_y$ and $-p_y\}$ states from Eq. (16.85) are occupied, and there is no free eigenstate until one goes up to the next Landau level which is a distance $\hbar\omega_c$ away in energy space. The Hall current in a full band is like a diamagnetic current which is associated with an equilibrium state of the system in magnetic field. If one introduced an obstacle in the path of the carriers, the current would flow around the obstacle, and its total value would remain the same (Prange 1981). Carriers cannot be scattered back and redistributed in other locations of space because all the eigenstates are full. If they could, then the eigenstates would be

redistributed or mixed into higher Landau bands, and the Fermi level would not remain in the mobility gap. The system creates new disturbed or scattered eigenstates which are linear combinations of the plane wave states $\exp\left(\frac{ip_y y}{\hbar}\right)$ considered above in Eq. (16.82). Every time $\{+p_y$ and $-p_y\}$ current contributions are added in this admixture, the net current is the same amount as for the undisturbed pairs. If the disturbance is such as to seriously mix the Landau bands, the Landau levels lose their identity, then the Fermi level no longer stays in a mobility gap, and the semiclassical conditions and Hall current are recovered. The same is true if the electron-phonon scattering destroys the quantum coherence of the Landau eigenstates making them so broad that they overlap each other.

16.1.11 The Fractional Quantum Hall Effect

Going back to the Hall experiment reveals that more detail appears in the Hall resistance structure as we go to higher and higher mobility and larger magnetic fields as shown in Fig. 16.18. The additional substructures are believed to be caused by

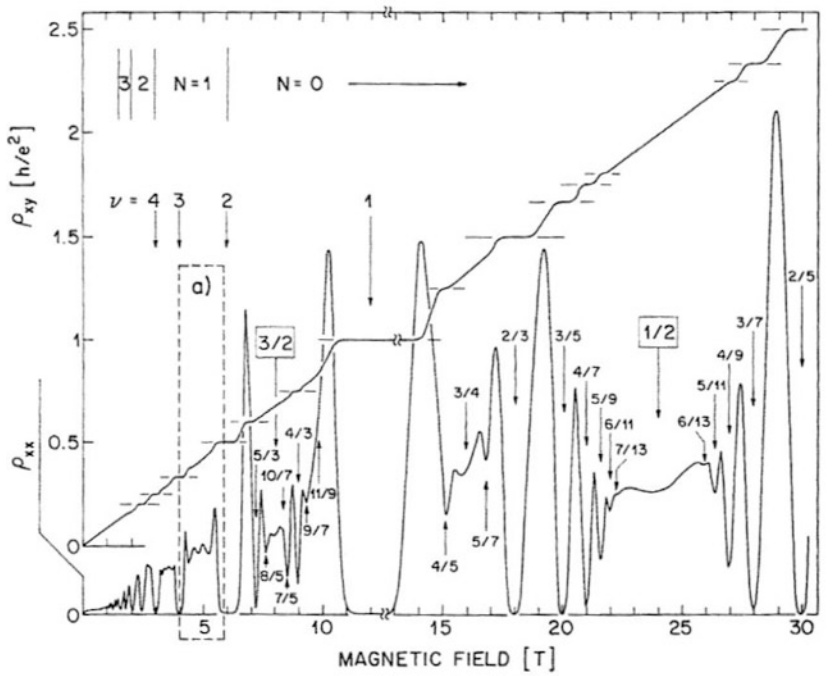


Fig. 16.18 Longitudinal resistivity ρ_{xx} and transverse resistivity ρ_{xy} of a high mobility two-dimensional electron ga at 150mK, showing the fractional quantum Hall effect filling factor ν is indicated and ρ_{xx} is reduced by a factor of 2.5 at high fields. (Reprinted with permission from Physical Review Letters Vol. 59, Willett, R., Eisenstein, H., Stoermer, H., Tsui, D., "Observation of an even denominator quantum number in the Fractional Quantum Hall effect," p. 1776, Fig. 1. Copyright 1987, American Physical Society)

electron-electron interactions. First we recall from Chap. 14 that in a 2DEG the screening is not as effective as in 3D. This is further accentuated by the high magnetic fields which produce gaps in the energy spectra and thus make it more difficult for charges to move and respond to the presence of other charges. In the high mobility 2DEG, we are dealing with mean-free paths of order 10 μm or more. The most serious disturbance seen by an electron in its "magnetic orbit," apart from the edge of the sample, is the presence of other electrons. In the presence of a strong magnetic field, the field caused by the other carrier constitutes a serious disturbance and generates novel forms of many-electron organizations, similar but even more complex than superconductivity. The charges and spins correlate their motion in pairs, threes, and even in lattices. . ., etc., in such a way as to reduce the total potential Coulomb energy. These coupled particles are also called quasi-particles and create their own energy gaps and elementary excitations. The excitations across the many body energy gaps can appear as particles with fractional charges, and this then produces the fractional quantum Hall effect (FQHE) characteristics shown in Fig. 16.18. One can see in this diagram that in addition to the usual Landau bandgaps, the system now exhibits new zeros in its "longitudinal resistivity" as defined by Eq. (16.90) and new plateaus in the "Hall resistivity" (Eq. (16.89)). At low densities, Wigner (1934) has shown that an electron gas prefers to crystallize on a lattice in order to minimize its potential energy. In a so-called Wigner crystal, electronic motion is then similar to phononic vibrations, and the system becomes an insulator. This trend is enhanced by a strong magnetic field which forces the electrons to move in Landau orbits. Now it can happen that for various electron densities or Landau band filling, "partial crystallization" takes place in a magnetic field, which is to say that the electron motions become correlated as described by the Laughlin wavefunction, yet they remain fluid and do not assume a rigid lattice-like correlation of a classic Wigner crystal. Each electron is not shared at a group of sites, and each site now appears to carry a fractional part of the charge. The new electronic "self-organization" is such that at any time, a fraction of the charge builds a Wigner type lattice to lower the Coulomb energy of the electron gas and is consequently immobilized, while the remaining fraction can conduct through the gaps of the Landau-Wigner lattice (fluid), giving the illusion of fractional charges (Laughlin 1983). Localized and mobile charges can exchange places in this highly correlated Wigner-Landau fluid. As in superconductivity, electrons add at one end (input electrode) into the many body collective state and leave at the other end (output) leaving the "many body collective" more or less intact. The charge transfer is in effect therefore the continual reorganization and exchange of particles with a many body state. The detailed discussion of this fascinating topic is however far beyond the scope of this book, and the reader is referred to the original literature on this subject (see Tsui et al. 1983 and Laughlin 1983).

16.1.12 Landau-Stark-Wannier States

The reader should note that we can also solve for the superlattice with an electric field in the growth direction as we did in Sect. 16.1.5 and also now allow a magnetic field in the growth direction (perpendicular to the plane). The combination of electric and magnetic field in the growth direction then generates the so-called Landau-Stark-Wannier bands. The wavefunctions are:

$$\Psi_{n,l,\nu}(p_y;x,y,z) = \left(\frac{1}{L_y}\right)^{1/2} \exp\left[\frac{i}{\hbar}(p_y y + p_z z)\right] \phi_n\left(x - x_{p_y}\right) \frac{1}{\sqrt{c}} J_{\frac{z}{c}-\nu}\left(\frac{2t_l}{qFc}\right) \tag{16.92}$$

$$\phi_n(x) = \left(\frac{1}{2^n n!\sqrt{\pi}}\right)^{1/2} \frac{1}{\sqrt{a_0}} \exp\left[-\frac{1}{2}(x/a_0)^2\right] H_n(x/a_0) \tag{16.93}$$

The energies are p_y, independent as in the 2DEG, and have the same degeneracy per Landau band:

$$E_{n,l,\nu} = \hbar\omega_c(n + 1/2) + qFc\nu + E_l^0 \tag{16.94}$$

where ν is again an integer which ranges from $[-\infty, \infty]$. When the magnetic field is applied in the plane of the superlattice, with an electric field in the growth direction as above, the mathematics is somewhat more complicated but still tractable in terms of the so-called Mathieu functions (Movaghar 1987). The electron moves in Landau orbits which are interrupted by the SL potential. The Landau confinement now competes with the Stark-Wannier localization, producing unusual electron dynamics. Novel effects are expected when the cyclotron radius $l_c = \sqrt{\frac{\hbar}{qB}}$ (see Eq. (10.119)) starts to compete with the Stark-Wannier localization length (Eq. (16.65)). This configuration has not been studied much because of the complexity of the experimental process involved. Applications could be envisaged in the fine-tuning of energy levels and performances of the quantum cascade laser (QCL). However fascinating the physics, applying a magnetic field of 5 T or more changes the size and cost of the device to such an extent that it renders it normally impractical as a "simple" laser system. But recently some workers have indeed made QCL structures which operate in perpendicular to the plane magnetic fields and which show promise as terahertz (THz) photodetectors and terahertz lasers (Scalari et al. 2006).

16.1.13 The Effective Mass of Carriers: Cyclotron Resonance

Using light it is possible to excite carriers from one Landau level to the other. The difference in energy between two adjacent subbands is $h\nu = \hbar\frac{qB}{m^*} \sim 10^{-4} B\frac{m}{m^*}$ eV where B is measured in Tesla. So an absorption experiment in the far infrared and

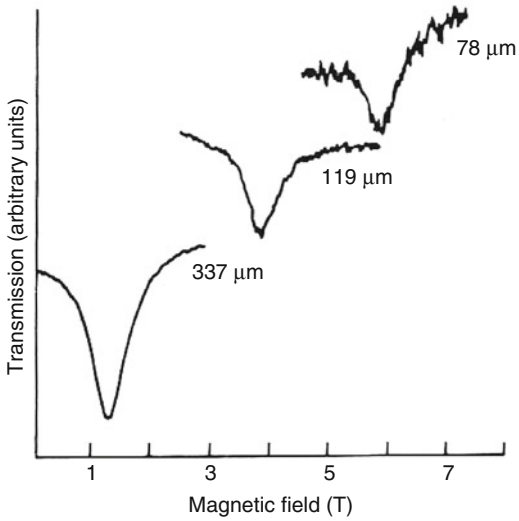


Fig. 16.19 Experimental recording of the transmission of 337, 119, and 78 μm radiation by $Ga_{0.47}In_{0.53}As$ as a function of magnetic field $\mu = 60000 cm^2/V/s$ (Copyright 1989 from "The MOCVD Challenge, Vol. 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications," Razeghi, M., p. 145, Fig. 3.65. Reproduced with permission of Routledge/Taylor & Francis Group, LLC)

at low temperatures should reveal distinct inter-Landau subband absorption processes, and the position of the energy resonance should allow us to deduce the effective mass. This is indeed one of the standard ways of measuring the effective mass, and an example is shown in Fig. 16.19. In materials with small effective mass, it is also possible to use this method to study how the effective mass changes with temperature and magnetic field. The way the change in the effective mass comes about can be seen from the Kane formula for the effective mass as discussed in Chap. 5. Temperature changes the bandgap; it reduces it in most semiconductors because the lattice tends to expand. The change in bandgap is roughly $\sim \gamma_T k_b T$ where the constant γ_T is of order 1 and varies from material to material. The change in bandgap will also according to Eq. (5.82) change the effective mass. However these averaged Kane effective mass corrections are very small, and one has to consider other corrections to the electronic energies as well. For example, one needs to include effects such as electron-phonon scattering and the resulting energy shifts which were briefly considered in Chap. 16. The excitation of phonons with temperature causes disorder, which scatters the electrons, and this gives rise to energy shifts which are corrections and appear as temperature-dependent energies, effective masses, and lifetimes. Though interesting and important, these corrections are second-order effects. They are important in low bandgap and small effective mass materials, such as InSb, but they constitute a specialized subject which is beyond the scope of this book.

16.1.14 Summary

In this chapter we introduced the reader to the way one would calculate electrical conduction using quantum mechanical methods. We used a simple example of a

structure one could engineer using modern growth techniques and worked it through. The generalization to more complex systems is in principle then straightforward. We explained how one can, when appropriate, relate quantum transport to the classical Drude method. We introduced also the double barrier system and explained how one can have a negative resistance and engineer the so-called negative differential resistance (NDR) device. The effect of a magnetic field perpendicular to the plane of motion in a two-dimensional electron gas (2DEG) was also considered. We showed how the Landau levels give rise to "Shubnikov de Haas" oscillations in the conductivity and the quantum Hall effect. Finally we studied what would happen to a carrier which is in a narrow band under the influence of an electric field. We derived the Stark-Wannier (SW) states and introduced the student to the fascinating new physics that this involves. This new physics is now realizable with semiconductor molecular beam epitaxy (MBE) and metalorganic chemical vapor deposition (MOCVD) growth techniques.

16.2 Electron-Phonon Interactions

16.2.1 Introduction

In Sect. 16.2.1 we treated electrical transport. There we made the observation that one of the reasons why carriers scatter and lose momentum is because they interact with the lattice vibrations, and this causes them to either gain or lose energy. Such processes, during which energy is exchanged, are called "inelastic scattering" processes. But electron-lattice interactions do much more than just cause resistance. They are also at the origin of such phenomena as superconductivity where phonons help two electrons (fermions) to bind together to form bosons (integer spin particles) which can condense into superfluids. But we will not go as far as that in this book and focus on some very elementary properties of electron-phonon coupling mainly the ones which enter transport properties.

Let us now look at the structure of the electron-lattice interaction, this coupling is also called the electron-phonon interaction. The reason for such a coupling is that when atoms move around, vibrate, and oscillate, they no longer form perfectly periodic arrays of potentials. But the electron Bloch functions were derived under the assumption that the system is at $T = 0$ and that the lattice is perfectly periodic. So these deviations from periodicity caused by thermal excitations cause a perturbation on the electron motion in Bloch bands; this causes them to scatter from one Bloch state k to another Bloch state k', exchanging momentum and energy in the process. The momentum difference is supplied by the lattice waves or, in quantized form, by the "phonons." One assumption which facilitates the analysis is the so-called Born-Oppenheimer approximation. This approximation is based on the observation that the electronic motion is very much faster than the lattice atom motion, the time scales being $\sim 10^{14}$ Hz (electron bandwidth) compared to 10^{12} Hz (lattice vibrational frequency at the Debye value), respectively. This implies that in most solids of interest to us, the electrons, when they move, see a more or less frozen lattice and

have time to build new eigenstates before the atom configurations change substantially. So the electrons form their new disturbed energy bands before the lattice has had time to change; this happens at least after the unit of time for the lattice, which is a lattice vibration. When the lattice configuration changes, the electrons respond by transferring to the new energy states formed by the new lattice configuration while conserving the total energy. This observation is also the basis for the methods used to perform numerical simulations of electron-phonon coupling in solids. One starts from a given lattice configuration, then solves for the electron energy levels, then allows a new configuration to evolve which obeys the lattice dynamics which can be assumed to obey Newton's laws. Then one solves again for the electron states, under the assumption that the new eigenstates obey trajectories which have the same total energy, vibrational and electronic. Numerical solutions of coupled electron-lattice systems are computationally very challenging and in their infancy. Fortunately, exact solutions are easier in nanoparticles and nanostructures where they are also more needed and much more relevant.

Now consider the formal derivation of the coupling with the help of which we can study the scattering rates using perturbation theory. We write for the electron ion interaction as usual:

$$H_{\text{el-ion}} = \sum_{l} V_{\text{el-ion}}\left(\vec{r} - \vec{R}_l\right) \tag{16.95}$$

Allowing the atomic variables $\vec{R}$, where $\vec{r}$ is the electron position, to deviate slightly from their equilibrium positions permits us to write the Hamiltonian in terms of one which describes the equilibrium periodic position and a nonperiodic term which scatters the electrons:

$$H_{\text{el-ion}} = H^0_{\text{el-ion}} + H_{\text{el-ph}} \tag{16.96}$$

Allowing the deviations from equilibrium of the α_{th} atom in the n_{th} Wigner-Seitz cell to be called $u_{n,\alpha}\left(\vec{R}\right)$, we can write:

$$V\left(\vec{r} - \vec{R}_{n\alpha} - \vec{u}_{n\alpha}\right) = V_\alpha\left(\vec{r} - \vec{R}_{n\alpha}\right) - \vec{u}_{n\alpha} \cdot \vec{\nabla}\, V_\alpha\left(\vec{r} - \vec{R}_{n\alpha}\right) \tag{16.97}$$

Now we note that the atomic deviation from equilibrium can be expanded in terms of the normal coordinates. The momentum index $\vec{q}$ should not be confused with the charge q:

$$\vec{u}_\alpha\left(\vec{R}\right) = \frac{1}{\sqrt{NM_\alpha}} \sum_{\vec{q}} Q_{\vec{q}}\, \vec{e}_{\alpha,\vec{q}}\, \exp\left[i\, \vec{q} \cdot \vec{R}\right] \tag{16.98}$$

where $\vec{e}_{\alpha,\vec{q}}$ is the polarization of the vibrational motion of the α_{th} atom in the q_{th} mode at the $\boldsymbol{R}$ site and M_s is the ionic masses. We can write the coupling as:

$$H_{\text{el-ph}} = -\sum_{\alpha,n} \frac{1}{\sqrt{NM_\alpha}} \sum_{\vec{q}} Q_{\vec{q}} \vec{e}_{\alpha,\vec{q}} \exp\left[i\,\vec{q}\cdot\vec{R}_n\right]\cdot\vec{\nabla}\, V_\alpha\left(\vec{r}-\vec{R}_{n\alpha}\right) \tag{16.99}$$

The adiabatic or Born-Oppenheimer approximation stipulates that one can write the total wavefunction as a product of electron and phonon wavefunctions so that:

$$\Psi_{\text{total}} = \Psi\left(\vec{r},\vec{R}\right)\Phi\left(\vec{R}\right) \tag{16.100}$$

where Φ is the wavefunction of all the ions and $\Psi\left(\vec{r},\vec{R}\right)$ is the wavefunction of the electrons in the instantaneous potential of the ions. When the ions move, the potential changes as they move, and we have a time evolving Hamiltonian where the electron potential at any time depends on the instantaneous position of the ions and trajectories conserve total energy. The motion of the ions was treated in Sect. 16.2.1 using classical mechanics. In principle the ions obey also the Schrödinger equation (see Eq. (4.8)) which can be written as:

$$H_{\text{ions}} = \sum_l \frac{p_l^2}{2M_l} + \sum_{l,m} K_{l,m}\left(\vec{R}_m - \vec{R}_l\right)\vec{u}_l\vec{u}_m + H_{\text{nonl}} \tag{16.101a}$$

where $\vec{u}_s$ are the displacements from equilibrium, p_l is the ionic momentum operator, $K_{l,m}$ is the restoring force per unit displacement, and H_{nonl} is the nonlinear term which is due to anharmonic forces (higher powers in u). It was shown in Chap. 6 that in the harmonic approximation, the lattice waves propagate as harmonic plane waves with solutions of the type:

$$\vec{u}\left(\vec{k},\omega\right) = \vec{u}_0 \exp\left[i\left(\vec{k}\cdot\vec{r}-\omega t\right)\right] \tag{16.101b}$$

with acoustic and optic branch frequency dispersions $\omega_{\vec{q},b}$ as shown in Chap. 6 and where b denotes the branch. We then assumed that in quantum mechanics, the lattice vibrations become quantized and can be thought of as particles with energies $\hbar\omega_{\vec{q},b}$, the so-called phonons. In quantum mechanics, and in the harmonic approximation, the solutions of Eq. (16.101) are the Bloch version of the localized harmonic oscillator wavefunctions.

In quantum mechanics, we note that the total wavefunction which describes the noninteracting unperturbed electron gas and the unperturbed phonon system, or lattice vibrations system, can be written as a product of the two wavefunctions:

$$\Psi = \Psi_{n\vec{k}}\left\{\phi_{\vec{q}_1},b\phi_{\vec{q}_2},b\ldots\ldots\phi_{\vec{q}_N},b\right\} \tag{16.102}$$

where $\psi_{\mathbf{nk}}$ is the electron state in the band index n and $\phi_{\vec{q},b}\left(\vec{R}_1,\vec{R}_2,\ldots\vec{R}_N\right)$ the wavefunction of a phonon with momentum index $\vec{q}$ in the b-branch (acoustic or

optic). We abbreviate the collective set of atomic displacements by the symbol $\vec{\Omega}$ so that the matrix element between two states ϕ, ϕ' which differ by the occupation of one phonon mode, abbreviated by n, becomes:

$$\begin{aligned}\int d\Omega_{\vec{q},b}\phi_{\vec{q}'',b}(\vec{Q})Q_{\vec{q},b'}\phi_{\vec{q}',b''}(\vec{Q}) = &\left(\frac{\hbar}{2\omega_{\vec{q},b}}\right)^{1/2}\delta_{\vec{q},\vec{q}',\vec{q}''}\delta_{b,b',b''}\\ &\times\left\{\left(n_{q,b}(\omega_{q,b})\right)^{1/2}\delta_{n',n-1}\right.\\ &\left.+\left(1+n_{q,b}(\omega_{q,b})\right)^{1/2}\delta_{n',n+1}\right.\end{aligned} \tag{16.103}$$

where the $n_{\vec{k},b}$ is the phonon occupation probability of the $(\vec{k},b)$ mode and for which the thermodynamic average is:

$$< n_{\vec{k},b} >= \frac{1}{e^{\hbar\omega_{\vec{k},b}/kT}-1} \tag{16.104}$$

It is convenient to use the shorthand Dirac notation to denote the phonon wavefunction (one branch only for simplicity):

$$\Psi_{n,\vec{k}}\Pi\phi_{q,b}\ldots\phi_{q,b} = \Psi_{n,\vec{k}}\left|n_{q_1,b}\vec{q}_1,b;n_{q_2,b}\vec{q}_2,b;\ldots;n_{q_N,b}\vec{q}_N,b\right\rangle \tag{16.105}$$

where there are $n_{q,\,b}$ phonons in the state $(q,b)\ldots$, etc. Of course one can use the Dirac notation for electrons too by replacing $\Psi_{n,\vec{k}} \rightarrow \left|(n,\vec{k})\right\rangle_e$ so that:

$$\Psi_{n,\vec{k}}\Pi\phi_{q,b}\ldots\phi_{q,b} = \left|n,\vec{k}\right\rangle_e\left|n_{q_1,b}\vec{q}_1,b;n_{q_2,b}\vec{q}_2,b;\ldots;n_{q_N,b}\vec{q}_N,b\right\rangle_{\text{ph}} \tag{16.106}$$

Then one can write the normal coordinates as operators which act on the Dirac state, and this is called the method of second quantization:

$$Q_{\vec{q},b} = \left(\frac{\hbar}{\omega_{q,b}}\right)^{1/2}\left(a^{+}_{-\vec{q},b}+a_{\vec{q},b}\right) \tag{16.107}$$

such that the operator $a_{\vec{q},b}$ removes a phonon from the state $(\vec{q},b)$ and $a^{+}_{\vec{q},b}$ adds a phonon $(\vec{q},b)$, for example:

$$\begin{aligned}&a^{+}_{q_s,b}\left|n_{q_1,b}\vec{q}_1,b;n_{q_2,b}\vec{q}_2,b;\ldots;n_{q_N,b}\vec{q}_N,b\right\rangle\\ &=\left|n_{q_1,b}\vec{q}_1,b;\left(n_{q_s,b}+1\right)\vec{q}_s,b;\ldots;n_{q_N,b}\vec{q}_N,b\right\rangle\end{aligned} \tag{16.108}$$

$$a_{q_s,b}\left|n_{q_1,b}\vec{q}_1,b;n_{q_2,b}\vec{q}_2,b;\ldots;n_{q_N,b}\vec{q}_N,b\right\rangle = \left|n_{q_1,b}\vec{q}_1,b;\left(n_{q_s,b}-1\right)\vec{q}_s,b;\ldots;n_{q_N,b}\vec{q}_N,b\right\rangle \tag{16.109}$$

Now the electron-phonon interaction can be written in the much more elegant form using the above creation and annihilation operators defined in Eq. (16.108) as:

$$H_{\text{el-ph}} = -\sum_{\alpha,n}\frac{1}{\sqrt{NM_\alpha}}\sum_{\vec{q}}\left(\frac{\hbar}{\omega_{q,b}}\right)^{1/2}\left[a^+_{-q,b}+a_{\vec{q},b}\right]\underset{\alpha,\vec{q}}{\vec{e}}\exp\left[i\vec{q}\cdot\underset{n}{\vec{R}}\right]\cdot\nabla V_\alpha\left(\vec{r}-\underset{n\alpha}{\vec{R}}\right) \tag{16.110}$$

When this operator acts on an electron-phonon basis state of the type (Eq. (15.108)), it raises and lowers the corresponding phonon occupation, and the r dependent terms act on the electron part of the wavefunction. The interaction with the longitudinal optical modes in a semiconductor are most important and take on the simple form (Note: q is charge):

$$H_{\text{el-ph,Lo}} = i\left\{\frac{q^2\hbar\omega_{Lo}}{2\varepsilon_0\Omega^{1/2}}\left[\frac{1}{\varepsilon(\infty)}-\frac{1}{\varepsilon(0)}\right]\right\}^{1/2}\sum_{\vec{k}}\left(\frac{1}{k}\right)\left[-a_k^+\exp\left[-i\vec{k}\cdot\vec{r}\right]+a_k\exp\left(i\vec{k}\cdot\vec{r}\right)\right] \tag{16.111}$$

where $\varepsilon(\infty)$, $\varepsilon(0)$ are the high and zero frequency permittivities and ω_{LO} the LO optical phonon frequency and Ω is the volume. The coupling with the acoustic mode gives:

$$H_{\text{el-ph, ac}} = \sum_{\vec{k}}\left(\frac{\hbar k}{2\Omega\rho_d v_s}\right)^{1/2} iD_{\text{ac}}\left[-a_k^+\exp\left(-i\vec{k}\cdot\vec{r}\right)+a_k\exp\left(i\vec{k}\cdot\vec{r}\right)\right] \tag{16.112}$$

where v_{s} is the velocity of sound, ρ_d the density, and D_{ac} is called the deformation potential.

Alternatively one can avoid the use of creation and annihilation operators and think of the electron-phonon interaction as a time-dependent perturbation of the lattice vibrations on the electron system, so that Eq. (16.65) is also:

$$H_{\text{el-ph, ac}} = \sum_{\vec{k}}\left(\frac{\hbar k}{2\Omega\rho_d v_s}\right)^{1/2} iD_{\text{ac}}\left[-e^{i\omega_{\vec{k}}t}\exp\left(-i\vec{k}\cdot\vec{r}\right)+e^{-i\omega_{\vec{k}}t}\exp\left(i\vec{k}\cdot\vec{r}\right)\right] \tag{16.113}$$

Consider the first-order scattering of Bloch electrons by acoustic phonons using the Fermi golden rule (Eq. (10.65)). The result is for an electron scattering process from k to k' with phonon emission:

$$\frac{1}{\tau_{\text{ac}}\left(\vec{k},\vec{k'}\right)} = \frac{2\pi}{\hbar}\sum_{\vec{k''}}\left(< n_{\vec{k''},\text{ac}} > +1\right) \times \frac{\hbar k''}{2\Omega\rho_d v_s} D_{\text{ac}}^2 \delta_{\vec{k}-\vec{k''},\vec{k'}} \delta\left(\varepsilon\left(\vec{k}-\vec{k''}\right) - \varepsilon\left(\vec{k}\right) + \hbar\omega_{\text{ac}}\left(\vec{k''}\right)\right) \tag{16.114}$$

Initial state was $\left|\vec{k}\right\rangle_{\text{el}} \left|n_{k''}\ldots\right\rangle_{\text{phon}}$ and final state $\left|\vec{k'}\right\rangle_{\text{el}} \left|n_{\vec{k''}}.+1,..\right\rangle_{\text{phon}}$. And for absorption we find:

$$\frac{1}{\tau_{\text{ac}}\left(\vec{k},\vec{k'}\right)} = \frac{2\pi}{\hbar}\sum_{\vec{k''}} < n_{\vec{k''},\text{ac}} > \frac{\hbar k''}{2\Omega\rho_d v_s} D_{\text{ac}}^2 \delta_{\vec{k}+\vec{k''},\vec{k'}} \delta\left(\varepsilon\left(\vec{k}+\vec{k''}\right) - \varepsilon\left(\vec{k}\right) - \hbar\omega_{\text{ac}}\left(\vec{k''}\right)\right) \tag{16.115}$$

Note that when evaluating the matrix elements, we have products and answers of the form ${}_{\text{phon}}\left\langle .n_{\vec{k}},\vec{k}\,..\left|a_{\vec{k}}^{+}a_{\vec{k}}\right|.\ldots.n_{\vec{k}},\vec{k}\right\rangle_{\text{phon}} = n_{\vec{k}}$ and ${}_{\text{phon}}\left\langle .n_{\vec{k}},\vec{k}\,..\left|a_{\vec{k}}a_{\vec{k}}^{+}\right|.\ldots.n_{\vec{k}},\vec{k}\right\rangle_{\text{phon}} = \left(1+n_{\vec{k}}\right)$

The sum over $\vec{q}$ can be transformed into an integral using:

$$\sum_{\vec{q}} \to \frac{\Omega}{(2\pi)^3}\int d\,\vec{q} \tag{16.116}$$

At high temperatures we also have the approximation:

$$< n_{\vec{k},b} > = \frac{1}{e^{\hbar\omega_{\vec{k},b}/kT} - 1} \sim < n_{\vec{k},b} > +1 \tilde{} kT/\hbar\omega_{\vec{k},b} \tag{16.117}$$

so that Eq. (16.105) reduce to the simple form:

$$\frac{1}{\tau_{\text{ac}}\left(\vec{K}\right)}_{\vec{k}-\vec{k'}=\vec{K}} = \frac{2\pi}{\hbar}\frac{k_B T}{\rho_d v_s^2} D_{\text{ac}}^2 g_V\left[\varepsilon\left(\vec{K}\right)\right] \tag{16.118}$$

where $g_V\left(\varepsilon\left(\vec{K}\right)\right)$ is the density of states per unit volume at energy $\varepsilon\left(\vec{K}\right)$. The scattering rate is straightforward to evaluate for optic modes too, and the formulae

can also be easily applied to allow for interband and intersubband scattering in quantum wells.

If we wish to avoid dealing with the creation and annihilation operators and phonon wavefunctions explicitly altogether and use the time-dependent perturbation form (Eq. (16.20)) of the interaction, then we have to put in the phonon probability factor by hand for phonon emission and absorption.

16.2.2 The Polaron Effective Mass and Energy

One of the novel feature introduced by the electron-phonon coupling is the polaron. The polaron is the electron surrounded by its polarization cloud. This polarization cloud is produced around it as a result of the coupling with the lattice. One can look at it this way: the electron has an electric field. This electric field will pull the positive ions toward the electron as it moves around. But moving the lattice ions generates a polarization, and this has dipole moment which couples to the electric field of the electron. The induced polarization couples with charge, and this produces a negative energy shift called polaron energy. As it moves, the electron drags this polarization cloud with itself, and this tends to make it look heavier. In other words, it acquires a polaron effective mass. To calculate this effect, one has to evaluate the second-order perturbation theory energy shift caused by, for example, the optic mode coupling given by Eq. (16.18). This is quite straightforward using Eq. (4.60) from Chap. 4, but it involves some integration. The result is that the electron acquires a new effective mass given by:

$$m^{**} = \frac{m^*}{1 - \alpha_{\text{ep}}/6} \tag{16.119}$$

where (q is the charge):

$$\alpha_{\text{ep}} = \frac{q^2}{8\pi\varepsilon_0\hbar\omega_{oL}}\left(\frac{2m^*\omega_{oL}}{\hbar}\right)^{1/2}\left[\frac{1}{\varepsilon(\infty)} - \frac{1}{\varepsilon(0)}\right] \tag{16.120}$$

The constant entering Eq. (16.110) is tabulated in data banks for semiconductors. Typically the value of α_{ep} is ~0.08 for InP and 0.015 for InSb but as high as ~2.4 and 6.6 for LiI and RbBr, respectively. In highly polar substances, the perturbation method is no longer valid and Eq. (16.10) is not meaningful.

For weak coupling, the other lattice modes also give terms of similar structures, so that the typical increase of the effective mass is never less than a few percent for most semiconductors. There is, as we have observed, also a concomitant energy shift (lowering) caused by the lattice polarization. The energy shift is of the form:

$$\Delta E_{\text{polaron}} \sim -\alpha_{\text{ep}}\hbar\omega_{\text{oL}} \tag{16.121}$$

Thus, typically, GaAs polaron shifts are ~3 meV. The polaron energy shift becomes more pronounced the more confined the electron eigenstate is. Thus in a quantum well or quantum dot, the energy lowering will be bigger, reaching values ~20 meV in GaAs compounds. A more confined electron spends its time visiting on average fewer sites and bonds, and thus has a greater effect on its environment than a highly delocalized charge.

16.2.3 Summary

In this chapter, which concludes the description of transport in solids, we allowed the lattice atoms to move and therefore to produce deviations from pure Bloch symmetry. The self-motion of the lattice produced by temperature, in other words, thermal excitation of phonons, then gives rise to electron-phonon interactions. We derived the optic and acoustic coupling and gave a simple example on how to calculate the electron-acoustic phonon scattering relaxation time. In most crystalline metals and semiconductors, it is the e-phonon scattering which limits the resistance of the material and is even invoked as an explanation of superconductivity. We also introduced the reader to the concept of the polaron and showed how to estimate energy and effective mass shifts in the presence of longitudinal optic mode coupling. We have here given the reader only a very brief description of this very important interaction mechanism and refer the reader to the specialized literature .

Problems for Quantum Transport

1. An electron beam is incident on a barrier where the barrier height is equal to the electron energy and is 8 eV. The transmission coefficient is given by $T = 10^{-3}$, what is the width of the barrier?
2. Calculate the transmission coefficient for the structure depicted in Fig. 16.1 but this time for a well with potential energy $\{-V\}$.
3. Calculate the expectation value of the velocity in the y-direction for an electron in a 2DEG (drop the z-part of the wavefunction) in the presence of a magnetic field perpendicular to the pane and an electric field in the x-direction. (See Eq. (16.82)). If the total number of electrons per unit area is Ns, what is the total current in the y-direction as a function of electric field and magnetic field?
4. Describe the various mechanisms that reduce the mobility in a 2DEG and how they affect the mobility in the different temperature regions. What would happen if one could suppress the optic phonon scattering?

Problems for Electron-Phonon Interactions

1. Why do lattice vibrations cause electrical resistance? In an electronic scattering process, what is the difference between phonon emission and phonon absorption? Using the electron-acoustic phonon interaction from Eq. (16.20) and the Fermi golden rule from Chap. 10, derive the scattering rates given by Eqs. (16.21) and (16.22).
2. Explain what is meant by the electron-phonon interaction. Taking the one-dimensional diatomic chain treated in Chap. 6 as an example, illustrate with a simple diagram the difference between the coupling of an electron to an acoustic and an optic vibration of the chain. It is helpful to think of the electronic bands as overlapping atomic wavefunctions, i.e., to think with the "tight-binding model" of the electronic band structure.
3. In analogy to Eqs. (16.21) and (16.22), write down an expression for the "Fermi golden rule" relaxation rate (see Sect. 16.2.1) of free electrons when they scatter with optic phonons. Consider both phonon emission and absorption. Take the electron wavefunctions to be plane waves.
4. What are the factors which influence the magnitude of the polaron energy shift in a solid (look at Eqs. (16.110) and (16.111))? What materials would you use to make strongly polaronic semiconductors?

References

Ando T, Fowler AB, Stern F (1982) Electronic properties of two-dimensional systems. Rev Mod Phys 54:437

Davies JH (2000) The physics of low dimensional semiconductors. Cambridge University Press, Cambridge

Faist J, Capasso F, Sivco DL, Sirtori C, Hutchinson AL, Cho AY (1994) Quantum cascade laser. Science 264:553

Laughlin RB (1983) The fractional quantum Hall effect. Phys Rev Lett 50:1395

Movaghar B (1987) Theory of high field transport in semiconductor superlattice structures. Semicond Sci Technol 2:185

Prange RE (1981) Quantized Hall resistance and the measurement of the fine-structure constant. Phys Rev B 23:4802

Razeghi M (1989) MOCVD challenge vol.1 a survey of GaInAsP-InP for photonic and electronic applications. Adam Hilger, Bristol

Razeghi M, Tardella A, Davis RA, Long AP, Kelly MJ (1987) Negative differential resistance at room temperature from resonant tunneling GaInAS/InP double barrier structures. IEEE Electron Lett 23:116

Scalari G, Graf M, Hofstetter D, Faist J, Beere JH, Ritchie D (2006) A Thz quantum cascade detector in a strong perpendicular magnetic field. Semicond Sci Technol 21:1743

Slivken S, Evans A, David J, Razeghi M (2002) High average power high duty cycle quantum cascade laser. Appl Phys Lett 81:4321

Tsui DC, Stoermer HL, Hwang JC, Brooks JS, Naughton MJ (1983) The fractional quantum Hall effect. Phys Rev B 28:2274

Wigner EP (1934) On the interaction of electrons in metals. Phys Rev 46:1002

Further Reading

Ahmed H (1986) An integration microfabrication system for low dimensional structures and devices. In: Kelly MJ, Weisbuch C (eds) The physics and fabrication of microstructures and microdevices. Springer, Berlin, pp 435–442

Asada M, Miyamoto Y, Suematsu Y (1986) Gain and the threshold of 3-dimensional quantum-box lasers. IEEE J Quantum Electron 22:1915–1921

Ashcroft NW, Mermin ND (1976) Solid state physics. Holt, Rinehart and Winston, New York

Bastard G (1988) Wave mechanics applied to semiconductor heterostructures. Halsted Press, New York

Beaumont SP (1992) Quantum wires and dots-defect related effects. Physica Scripta 1992 (T45):196–199

Bockrath M, Cobden DH, Lu J, Rinzler AG, Smalley RE, Balents L, McEuen PL (1999) Luttinger liquid behaviour in carbon nanotubes. Lett Nat 397:598–601

Chuang SL (1995) Physics of optoelectronic devices. Wiley, New York

Cochran W (1973) The dynamics of atoms in crystals. Edward Arnold Limited, London

Cohen MM (1972) Introduction to the quantum theory of semiconductors. Gordon and Breach, New York

Dingle R (1975) Confined carrier quantum states in ultrathin semiconductor heterostructures. In: Queisser HJ (ed) Feskoerperprobleme XV. Springer, pp 21–48

Dingle R, Wiegmann W, Henry CH (1974) Quantum states of confined carriers in very thin $Al_xGa_{1-x}As$-GaAs- $Al_xGa_{1-x}As$ heterostructures. Phys Rev Lett 33:827–830

Einspruch NG, Frensley WR (1994) Heterostructures and quantum devices. Academic Press Limited, London

Ferry DK (1991) Semiconductors. Macmillan, New York

Hasko DG, Potts A, Cleaver JR, Smith C, Ahmed H (1988) Fabrication of sub-micrometer free standing single crystal GaAs and Si structures for quantum transport studies. J Vac Sci Technol B 6:1849–1851

Ibach H, Lüth H (1990) Solid-state physics: an introduction to theory and experiment. Springer, New York

Kasap SO (1997) Principles of engineering materials and devices. McGraw-Hill, New York

Kelly MJ (1995) Low-dimensional semiconductors: materials, physics, technology, devices. Oxford University Press, New York

Kittel C (1976) Introduction to solid state physics. Wiley, New York

Maxwell JC (1952) Matter and motion. Dover, New York

Peyghambarian N, Koch SW, Mysyrowicz A (1993) Introduction to semiconductor optics. Prentice-Hall, Englewood Cliffs

Razeghi M, Duchemin JP, Portal JC, Dmowski L, Remeni G, Nicolas RJ, Briggs A (1986) First observation of the quantum Hall effect in a $Ga_{0.47}In_{0.53}As$-InP heterostructure with three electric subbands. Appl Phys Lett 48:712–715

Reissland JA (1973) Physics of phonons. Wiley, London

Rosencher E, Vinter B (2002) Optoelectronics. Cambridge University Press, Cambridge

Sapoval B, Hermann C (1995) Physics of semiconductors. Springer, New York

Scherer A, Jewell J, Lee YH, Harbison J, Florez LT (1989) Fabrication of microlasers and microresonator optical switches. Appl Phys Lett 55:2724–2726

Sze S (1981) Physics of semiconductor devices, 2nd edn. Wiley, New York

Tewordt M, Law V, Kelly M, Newbury R, Pepper M, Peacock C (1990) Direct experimental determination of the tunneling time and transmission probability of electrons through a resonant tunneling system. J Phys Condens Matter 2:896–899

Weiner JS, Miller DAB, Chemla DJ, Damen TC, Burrus CA, Wood TH, Gossard AC, Wiegmann W (1985) Strong polarization sensitive electroabsorption in GaAs/AlGaAs quantum well waveguides. Appl Phys Lett 47:1148–1151

Weisbuch C, Vinter B (1991) Quantum semiconductor structures fundamentals and applications. Academic Press Inc Harcourt Brace Jovanovich, Publishers, Boston

Compound Semiconductors and Crystal Growth Techniques

17

17.1 Introduction

A key component in semiconductor microtechnology is the production and quality control of the basic semiconductor materials from which devices and integrated circuits are made. These semiconductor materials are usually composed of single crystals of high perfection and high purity.

Today, silicon technology has reached the stage where complex integrated circuits containing millions of transistors can be manufactured reproducibly and reliably. This is not only a result of the development of device technology but also the improvement of base material quality. For example, the silicon material that is now used for devices has an impurity concentration less than one part in ten billion. Unlike silicon, compound semiconductors consist of at least two different types of atoms. Compound semiconductors are emerging as important materials suitable for optoelectronic applications, which involve the optical and electrical properties of the semiconductors. Gallium arsenide is an example of a compound semiconductor material. Although its technology is not yet as mature as the one of silicon, there is currently much effort being done in order to achieve a very high circuit operational speed as a consequence of the high electron mobility in this material.

When improving the technology for a particular semiconductor material, a specific range of issues must be resolved before high-performance devices can be fabricated with a high degree of reproducibility and reliability. Only then can large-scale production be contemplated. An important consideration in this process, which will decide whether a material or technology will be commercially used, is the costs of implementation and production. To establish a new material technology or fabrication technique, it is essential to demonstrate that a significantly improved performance, lower costs, and/or new device functionalities will result.

In this chapter, we will first review the properties of major III-V compound semiconductors. We will then describe the current techniques used in the synthesis of semiconductor crystals. These are divided into two categories: single crystal

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_17

growth techniques and epitaxial growth techniques. The former is used to fabricate semiconductor crystals of macroscopic size that will be processed into substrates, while the latter is used to deposit thin films of a few micrometers (or less) onto one of these substrates.

17.2 III-V Semiconductor Alloys

17.2.1 III-V Binary Compounds

III-V binary semiconductors are compounds which involve one element from the group III and one from the group V columns of the periodic table. Table 17.1 lists some of the fundamental physical parameters of common binary III-V compounds. These binary compounds are the simplest III-V compounds and constitute the basis for more complex ternary or quaternary compounds.

17.2.2 III-V Ternary Compounds

When one additional element from the group III or group V is present and is distributed randomly in the crystal lattice, III_x-III_{1-x}-V or III-V_y-V_{1-y} ternary alloys can be achieved, where x and y are indices with values between 0 and 1. This allows to modify the alloy bandgap energy and lattice parameter.

The bandgap energy $E_g(x)$ of a ternary compound varies with the composition x as follows:

$$E_g(x) = E_g(0) + bx + cx^2 \tag{17.1}$$

where $E_g(0)$ is the bandgap energy of the binary compound corresponding to $x = 0$ and c is called the bowing parameter. The compositional dependence of the bandgap energy of various III-V ternary alloys at 300 K is given in Table 17.2 (Casey and Panish 1978).

The bowing parameter c can be theoretically determined (Van Vechten and Bergstresser 1970). It is especially helpful to estimate c when experimental data are unavailable.

The lattice constant a of ternary compounds can be calculated using Vegard's law. According to Vegard's law, the lattice constant of the ternary alloy $A_xB_{1-x}C$ can be expressed as follows:

$$a_{A_xB_{1-x}C} = xa_{AC} + (1-x)a_{BC} \tag{17.2}$$

where a_{AC} and a_{BC} are the lattice constants of the binary alloys AC and BC. Vegard's law is obeyed quite well in most of the III-V ternary alloys.

Table 17.1 Physical constants of some III-V binary compounds at 300 K

III-V binary compound	Average atomic number ($\bar{z}$)	Lattice parameter (Å)	Bandgap energy (eV)	Refractive index $\bar{n}$	Effective mass (m_e/m_0)	Effective mass (m_{hh}/m_0)	Effective mass (m_{lh}/m_0)	Dielectric constant ($\varepsilon/\varepsilon_0$)	Electron affinity χ (eV)
InSb	50	6.47937	0.17	4.0	0.0145	0.44	0.016	17.7	4.69
InAs	41	6.0584	0.36	3.520	0.022	0.41	0.025	14.6	4.45
GaSb	41	6.09593	0.73	3.820	0.044	0.33	0.056	15.7	4.03
InP	32	5.86875	1.35	3.450	0.078	0.8	0.012	12.4	4.4
GaAs	32	5.65321	1.424	3.655	0.065	0.45	0.082	13.1	4.5
A1Sb	32	6.1335	1.58	3.400	0.39	0.5	0.11	14.4	3.64
InN	28	$a = 3.545$ $c = 5.703$	1.9	2.56	0.11	0.5	0.17	–	–
A1As	23	5.6622	2.16	3.178	0.11		0.22	10.1	–
GaP	23	5.45117	2.26	3.452	0.35	0.86	0.14	11.1	4.0
GaN	19	$a = 3.189$ $c = 5.186$	3.44	2.35	0.2	0.8	–	10.4	–
A1P	14	5.451	2.45	3.027	–	0.63	0.20	–	–
A1N	10	$a = 3.112$ $c = 4.982$	6.2	2.2	–	–	–	9.14	–

Table 17.2 Compositional dependence of the bandgap energy in some III-V ternary compound semiconductors at 300 K (Casey and Panish 1978)

Ternary	Direct bandgap energy E_g (eV)
$Al_xGa_{1-x}As$	$E_g(x) = 1.424 + 1.247x$
$Al_xIn_{1-x}As$	$E_g(x) = 0.360 + 2.012x + 0.698x^2$
$Al_xGa_{1-x}Sb$	$E_g(x) = 0.726 + 1.139x + 0.368x^2$
$Al_xIn_{1-x}Sb$	$E_g(x) = 0.172 + 1.621x + 0.43x^2$
$Ga_xIn_{1-x}P$	$E_g(x) = 1.351 + 0.643x + 0.786x^2$
$Ga_xIn_{1-x}As$	$E_g(x) = 0.360 + 1.064x$
$Ga_xIn_{1-x}Sb$	$E_g(x) = 0.172 + 0.139x + 0.145x^2$
GaP_xAs_{1-x}	$E_g(x) = 1.424 + 1.15x + 0.176x^2$
$GaAs_xSb_{1-x}$	$E_g(x) = 0.726 - 0.502x + 1.2x^2$
InP_xAs_{1-x}	$E_g(x) = 0.36 + 0.891x + 0.101x^2$
$InAs_xSb_{1-x}$	$E_g(x) = 0.18 - 0.41x + 0.58x^2$

17.2.3 III-V Quaternary Compounds

Similarly, quaternary compounds can be obtained when there is a total of four different elements from the group III or group V columns distributed uniformly in the crystal lattice. The interest in these quaternary compounds has centered on their use in conjunction with binary and ternary alloys to form lattice-matched heterojunction structures with different bandgaps. Indeed, by controlling the composition of a quaternary alloy, it is possible to change both its bandgap energy and its lattice parameter. For example, the reduction of stress in $Al_xGa_{1-x}As$ layers grown on GaAs substrates can be done by introducing small amounts of P to realize the quaternary $Al_xGa_{1-x}P_yAs_{1-y}$. The $InP/Al_xGa_{1-x}P_yAs_{1-y}$ heterojunction serves as a successful example of a binary-quaternary lattice-matched system.

Ilegems and Panish (1974) calculated quaternary phase diagrams with the solid decomposed into ternary alloys: ABC, ACD, ABD, and BCD (where A and B are group III elements and C and D are group V elements). Jordan and Ilegems (1974) obtained equivalent formulations considering the solid as a mixture of binary alloys: AC, AD, BC, and BD. Assuming a linear dependence on composition of lattice parameter a_{AC} for the binary AC, and similarly for the other lattice parameters, the lattice parameter of the alloy $A_xB_{1-x}C_yD_{1-y}$ is:

$$a_{A_xB_{1-x}C_yD_{1-y}} = xya_{AC} + x(1-y)a_{AD} + (1-x)ya_{BC} + (1-x)(1-y)a_{BD} \quad (17.3)$$

The quaternary III-V alloys which can be used for multilayer heterostructures are listed in Table 17.3 along with the binary compounds to which they are lattice-matched.

The determination of the bandgap energy is more complicated. However, if the bowing parameter c is neglected, the bandgap energy may be approximated from that of the binaries, assuming a linear dependence:

$$E_g = xyE_{AC} + x(1-y)E_{AD} + (1-x)yE_{BC} + (1-x)(1-y)E_{BD} \quad (17.4)$$

Table 17.3 Binary to quaternary III-V lattice-matched systems of multilayer heterostructures (Casey and Panish 1978)

Quaternary	Lattice-matched binary	Wavelength, λ (μm)
$Al_xGa_{1-x}P_yAs_{1-y}$	GaAs	0.8–0.9
$Al_xGa_{1-x}As_ySb_{1-y}$	InP	1
$Al_xGa_{1-x}As_ySb_{1-y}$	InAs	3
$Al_xGa_{1-x}As_ySb_{1-y}$	GaSb	1.7
$Ga_xIn_{1-x}P_yAs_{1-y}$	GaAs, InP	1–1.7
$Ga_xIn_{1-x}P_ySb_{1-y}$	InP, GaSb, AlSb	2
$In(P_xAs_{1-x})_ySb_{1-y}$	AlSb, GaSb, InAs	2–4
$(Al_xGa_{1-x})_yIn_{1-y}P$	GaAs, $Al_xGa_{1-x}As$	0.57
$(Al_xGa_{1-x})_yIn_{1-y}As$	InP	0.8–1.5
$(Al_xGa_{1-x})_yIn_{1-y}Sb$	AlSb	1.1–2.1

By using advanced epitaxial growth techniques, such as the ones discussed in Sect. 17.5, multilayer structures of compounds with different bandgap-associated wavelengths can be synthesized.

Figure 17.1 is an illustration of the phase diagram for the GaInPAs-AlAs-AlP system, providing the bandgap energy and lattice parameters of the common ternary and quaternary III-V alloys. Each of the four corners of the central square corresponds to a binary III-V semiconductor. Each side of the square represents a III-III-V ternary alloy, such as $Ga_xIn_{1-x}P$ (bottom) and $Ga_xIn_{1-x}As$ (top), or a III-V-V ternary alloy, such as $GaP_{1-y}As_y$ (left) and $InP_{1-y}As_y$ (right). By selecting the composition of the different materials, it is possible to change their bandgap and therefore vary the optical properties of the semiconductor materials.

The inner part of the diagram corresponds to the quaternary$Ga_xIn_{1-x}P_{1-y}As_y$ compound. The curved lines indicate compounds with equal bandgap energy, and the solid lines represent those with equal lattice constants. By continuously varying the concentration of gallium, indium, phosphorus, and arsenic, one can vary the characteristics of $Ga_xIn_{1-x}P_{1-y}As_y$ in the range between those of indium arsenide (InAs), indium phosphide (InP), gallium arsenide (GaAs), and gallium phosphide (GaP) as shown in Fig. 17.1. Such formation of ternary and quaternary compounds enables the development of heterostructures, which have become essential for the design of high-performance electronic and optoelectronic devices, especially in semiconductor lasers.

For optoelectronic applications, two possible systems are of interest. One consists of compounds which are lattice-matched to GaAs substrate and their bandgap energy from 1.42 eV to 1.92 eV. These compounds are located on the thick solid line that begins from the upper left-hand corner and extends to the bottom of the $Ga_xIn_{1-x}P$ ternary edge. The second system consists of compounds lattice-matched to InP substrate and their bandgap energy between 0.75 eV to 1.35 eV.

The bandgap energy and lattice parameter of common II-VI, III-V, and IV-IV semiconductors can be easily represented in the diagram shown in Fig. 17.2. The

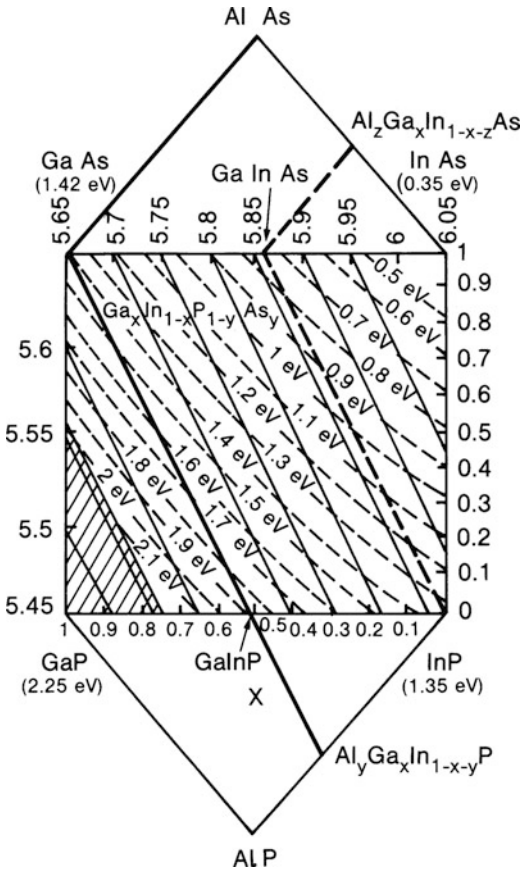


Fig. 17.1 The *x-y-z* compositional plane for quaternaries III-V alloys at 300 K. The solid lines in the square center region represent the *x-y* coordinates for which the quaternary alloy has a constant lattice parameter, while the curved dashed lines represent the *x-y* coordinates for which the alloy has a constant bandgap energy. The bold straight solid line represents the *x-y* coordinates for the quaternary alloys with the same lattice constant as GaAs. The bold straight dashed line represents the *x-y-z* coordinates for the quaternary alloys with the same lattice constant as InP (Copyright 1989 from The MOCVD challenge vol, 1: a survey of GaInAsP-InP for photonic and electronic applications. P. 4, Fig. 1.1 Reproduced by permission of Routledge/Taylor & Francis Group, LLC)

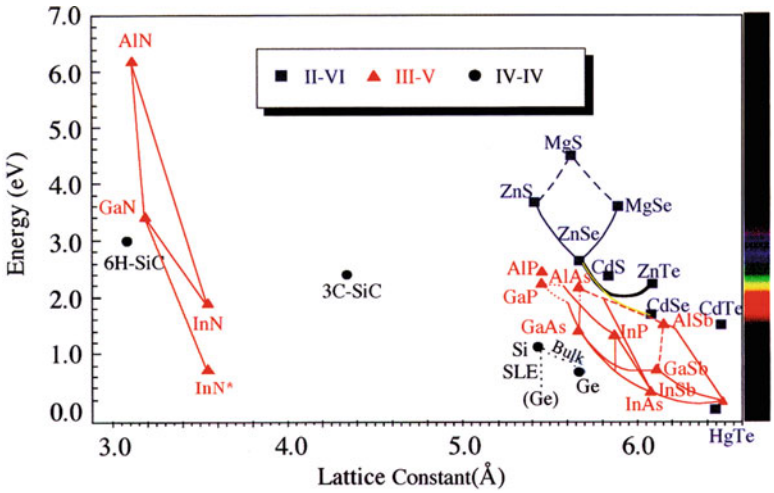


Fig. 17.2 Bandgap energy vs. lattice constant diagram of common semiconductors. A dashed line indicates an indirect bandgap material. (The bandgap energy of pure InN has been found to be 0.7 eV which is much smaller than the previously reported value of 1.9 eV)

lines connecting two compounds in the diagram correspond to the bandgap energy and lattice constant positions of ternary compounds involving the two binary semiconductor endpoints.

17.3 II-VI Compound Semiconductors

Right after the limitations of the elemental group IV semiconductors were exposed several decades ago, researchers started to study III-V and II-VI semiconductors more vigorously. Although not as popular as III-V compounds, II-VI semiconductors have been the focus of many intensive studies in the past few decades. One of the interesting properties of II-VI compounds is their direct energy gaps (with the exception of semi-metals: HgTe and HgSe), which is suitable for optoelectronic device applications. Perhaps the most celebrated II-VI optoelectronic devices are HgCdTe-based infrared photodetectors and focal plane arrays. Albeit facing recent challenges from III-V-based structures, these photodetectors are still the best choice especially in the near-infrared and mid-infrared range. In addition to photodetectors, visible light-emitting devices based on ZnSSe/ZnCdSe semiconductors have also been demonstrated in the II-VI material system.

As mentioned before, the II-VI family not only involves semiconductors but also a couple of semi-metallic compounds. For instance, HgTe is a semi-metal, while CdTe is a semiconductor with a bandgap energy of 1.6 eV. For the ternary HgCdTe compound, the bandgap energy ranges from 0 to 1.6 eV, depending on the Hg (or Cd) molar fraction. Table 17.4 lists some of the II-VI compound semiconductors and their respective bandgap energies, their crystalline structures, and the rate of change of their bandgap energy as a function of temperature (Ray 1969).

Table 17.4 Bandgap energy, crystal structure (W = Wurtzite, ZB = Zinc blende), and temperature coefficient (rate of change of bandgap energy as a function of temperature) for a few II-VI compounds. Ray (1969), Roberts and Zallen (1971)

Compound	E_g (eV)/Structure	dE_g/dT (10^{-4} eV/K)
ZnO	3.44/W	−9.5
ZnS	3.91/W, 3.84/ZB	−8.5, −4.6
CdS	2.58/W	−5.2
HgS	2.10/ZB	−9.0
ZnSe	2.80/W, 2.83/ZB	−8.0 (ZB)
CdSe	1.84/W	−4.6
HgSe	−0.1/ZB	–
ZnTe	2.39/ZB	−5.0
CdTe	1.60/ZB	−2.3
HgTe	−0.1/ZB	–

17.4 Bulk Single Crystal Growth Techniques

The starting point for virtually all semiconductor devices is in the form of flat template, known as the substrate which is made entirely of a single material. Its crucial features are that it is one single crystal across its entirety with no grain boundaries. The process of creating single crystal wafers is simpler if they are made purely from a single element, such as silicon. Elemental silicon is obtained by chemical decomposition of compounds such as $SiCl_4$ and SiH_4. Then the initial purification processes are performed, and the material is melted and cast into ingots. Upon cooling, careful control of the boundary between the molten material and solid is required; otherwise the material will be polycrystalline. Today, three methods have been developed to produce bulk single crystals for the epitaxial growth of most semiconductors: the Czochralski, Bridgman, and float-zone methods. A fourth technique, the Lely growth method, was also developed in order to produce substrates when a melt was not available. All of these methods will be discussed in the following subsections.

17.4.1 Czochralski Growth Method

The Czochralski (CZ) crystal growth method uses a quartz (SiO_2) crucible of high purity in which pieces of polycrystalline material, termed "charge," are heated above their melting point (e.g., 1415 °C for silicon). The crucible, shown in Fig. 17.3, is

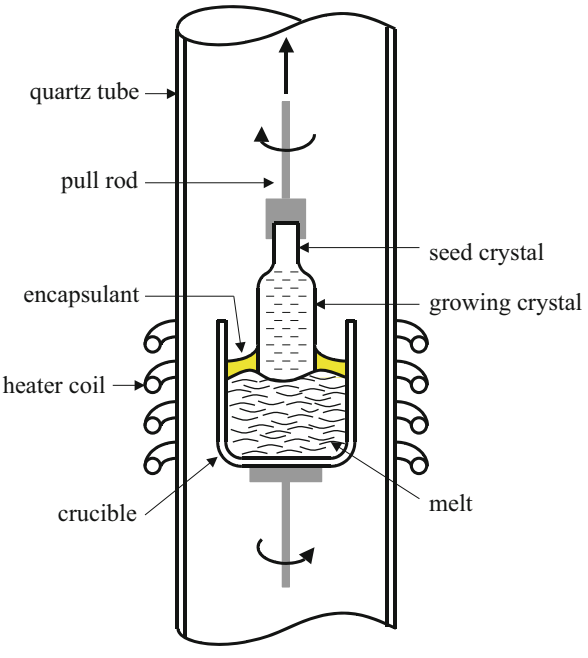


Fig. 17.3 Cross section of a furnace used for the growth of single crystal semiconductor boules by the Czochralski process, in which a tiny single crystal is suspended in a pool of hot molten material and is slowly drawn upward as the crystal grows from the melt. The resulting boule can have a diameter over 30 cm and a length up to 2 m

heated by either induction using radio-frequency (RF) energy or thermal resistance methods. A "seed" crystal, which is about 0.5 cm in diameter and 10 cm long, with the desired orientation is lowered into molten crystal, termed "melt," and then drawn up at a carefully controlled rate.

When the procedure is properly done, the material in the melt will make a transition into a solid-phase crystal at the solid-liquid interface, so the newly created material accurately replicates the crystal structure of the seed crystal (Fig. 17.3). The resulting single crystal is called the boule. Modern boules of silicon can reach diameters over 300 mm and lengths up to two meters. The Czochralski method is by far the most popular method, accounting for between 80 and 90% of all silicon crystals grown for the semiconductor industry.

Since both the molten semiconductor and the solid are at the same pressure and have approximately the same composition, crystallization results due to a reduction in temperature. As the melt is drawn up, it loses heat via radiation and convection to the inert gas. This heat loss results in a substantial thermal gradient across the liquid and solid interface. At this interface, additional energy must be lost to accommodate the latent heat of fusion of the solid. A control volume one-dimensional (in the x-axis) energy balance for the interface yields the following relation:

$$\left(-k_{\mathrm{l}}A\frac{dT}{dx}\bigg|_{\mathrm{l}}\right)-\left(-k_{\mathrm{s}}A\frac{dT}{dx}\bigg|_{\mathrm{s}}\right)=L\frac{dm}{dt} \tag{17.5}$$

where k_{l} and k_{s} are the thermal conductivity of the liquid and solid silicon at the melting point, respectively, A is the cross-sectional area of the boule, T the temperature, L the latent heat of fusion (~340 cal/g for silicon), and m is the mass of the growing solid silicon. Under normal conditions used for CZ growth, the heat diffusion from the liquid is small compared to the heat diffusion from the solid. This allows the equation above to be simplified and yields the following expression for the maximum velocity at which the solid can be pulled:

$$v_{\max}=\frac{kA\,dT}{L\,dm}=\frac{k\quad dT}{M_{\mathrm{V}}L\,dx}\bigg|_{\mathrm{s}} \tag{17.6}$$

where M_{V} is the solid density of the growing crystal. If the crystal is pulled with a velocity $v > v_{\max}$, then the solid cannot conduct enough heat away, and the material will not solidify in a single crystal. In practice, the pull rate of the seed crystal varies during the growth cycle. It is faster when growing the relatively narrow neck (5–12 inches per hour) so the generation of defects known as dislocations is minimized. Once the neck has been formed, the pull rate is reduced to form the shoulder of the crystal, finally approaching 2–4 inches per hour during the growth of the crystal body.

During the entire growth, the crucible rotates in one direction at 12–14 rotations per minute (rpm), while the seed holder rotates in the opposite direction at 6–8 rpm. This constant stirring prevents the formation of local hot or cold regions. The crystal diameter is monitored by an optical pyrometer which is focused at the interface

between the edge of the crystal and the melt. An automatic diameter control system maintains the correct crystal diameter through a feedback loop control. Argon is often used as the ambient gas during this crystal-pulling process. By carefully controlling the pull rate, the temperature of the crucible, and the rotation speed of both the crucible and the rod holding the seed, a precise control of the diameter of the crystal is obtained.

During the Czochralski growth process, several impurities will incorporate into the crystal. Since the crucibles are made of fused silica (SiO_2) and the growth process takes place at temperatures around 1500 °C, small amounts of oxygen will be incorporated into the boule. In order to reduce the concentration of oxygen impurities, the boule is usually grown under magnetic confinement. In this situation, a large magnetic field is directed perpendicularly to the pull direction, generating a Lorentz force. This force changes the motion of the ionized impurities in the melt so as to keep them away from the liquid/solid interface and therefore decrease the impurity concentration. Using this arrangement, the oxygen impurity concentration can be reduced from about 20 parts per million (ppm) to as low as 2 ppm.

It is also common to introduce dopant atoms into the melt in order to tailor the electrical properties of the final crystal, i.e., carrier type and concentration. Simply weighing the melt and introducing a proportional amount of impurity atoms is all that is theoretically required to control the carrier concentration. However, impurities tend to segregate at the liquid/solid interface, rather than being uniformly distributed inside the melt. This will in turn affect the amount of dopant incorporated into the growing solid. This behavior can be quantitatively characterized by a dimensionless parameter called the segregation constant k defined by:

$$k = \frac{C_s}{C_l} \tag{17.7}$$

where C_l and C_s are the impurity concentrations in the liquid and solid sides of the liquid/solid interface, respectively. Table 17.5 lists the values of the segregation constant for some common impurities in silicon.

Let us consider for example the case where $k > 1$. By definition, the concentration of impurity in the solid is greater than that in the melt. Therefore the impurity concentration in the melt decreases as the boule is pulled. The resulting crystal impurity concentration, C_s, can be expressed mathematically as:

Table 17.5 Segregation constants for a few common impurities in silicon

Impurity	k
Al	0.002
As	0.3
B	0.8
O	0.25
P	0.35
Sb	0.023

$$C_s = kC_o(1 - X)^{k-1} \tag{17.8}$$

where C_0 is the original impurity concentration and X is the fraction of the melt that has solidified.

The growth of GaAs with the Czochralski method is far more difficult than for silicon because of the vast difference between the vapor pressures of the constituents at the growth temperature of ~1250 °C: 0.0001 atm for gallium and 10,000 atm for arsenic. Liquid Encapsulated Czochralski (LEC) utilizes a tightly fitting disk and sealant around the melt chamber to prevent the out-diffusion of arsenic from the melt. The most commonly used sealant is boric oxide (B_2O_3). Additionally, pyrolytic boron nitride (pBN) crucibles are used instead of quartz (silicon oxide) in order to avoid silicon doping of the GaAs boule. Once the charge is molten, the seed crystal can be lowered through the boric oxide until it contacts the charge at which point it may be pulled.

Since the thermal conductivity of GaAs is about one-third that of silicon, the GaAs boule is not able to dissipate the latent heat of fusion as readily as silicon. Furthermore, the shear stress required to generate a dislocation in GaAs at the melting point is about one-fourth that in silicon. Consequently, the poorer thermal and mechanical properties allow GaAs boules to be only about 8 inches in diameter, and they contain many orders of magnitude larger defect densities than realized in silicon.

17.4.2 Bridgman Growth Method

The Bridgman crystal growth method is similar to the CZ method except for the fact that the material is completely kept inside the crucible during the entire heating and cooling processes, as shown in Fig. 17.4.

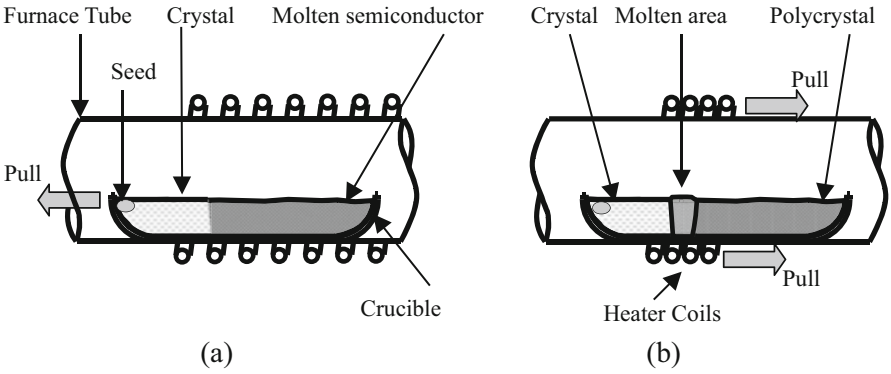


Fig. 17.4 Bridgman growth method in a crucible (**a**) solidification from one end of the melt (**b**) melting and solidification in a moving heated zone

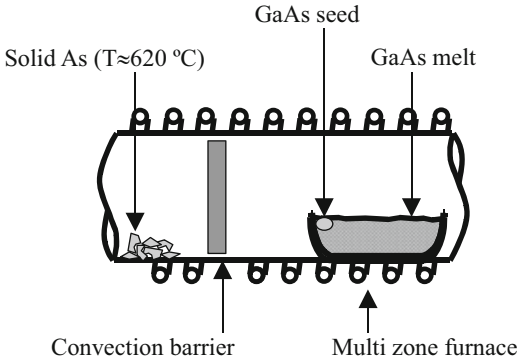


Fig. 17.5 Schematic diagram of the Bridgman growth method for a compound semiconductor such as gallium arsenide

A quartz crucible filled with material is pulled horizontally through a furnace tube. As the crucible is drawn slowly from the heated region into a colder region, the seed crystal induces single crystal growth. The shape of the resulting crystal is determined by that of the crucible. In a variation of this procedure, the heater may move instead of the crucible.

There are a couple of disadvantages associated with the Bridgman growth method which result from the fact that the material is constantly in contact with the crucible. First, the crucible wall introduces stresses in the solidifying semiconductor. These stresses will result in deviations from the perfect crystal structure. Also, at the high temperatures required for bulk crystal growth, silicon tends to adhere to the crucible.

In the case of compound semiconductors, the process is slightly different from that for silicon. The basic process is shown in Fig. 17.5 for gallium arsenide. The solid gallium and arsenic components are loaded into a fused silica ampoule which is then sealed. The arsenic in the chamber provides the overpressure necessary to maintain stoichiometry. A tube furnace is then slowly pulled past the charge. The temperature of the furnace is set to melt the charge when it is completely inside. As the furnace is pulled past the ampoule, the molten GaAs charge in the bottom of the ampoule recrystallizes. A seed crystal may be mounted so as to contact the melt.

Typical compound semiconductor boules grown by the Bridgman method have diameters of 2 inches. The growth of larger crystals requires very accurate control of the stoichiometry and the radial and axial temperature gradients. Dislocation densities of lower than 10^3 cm^{-2}, compared to 10^4 cm^{-2} for boules grown by CZ, are routinely achieved by using the Bridgman method. Roughly 75% of the compound semiconductor boules are grown by the Bridgman growth method.

17.4.3 Float-Zone Crystal Growth Method

The float-zone (FZ) crystal growth proceeds directly from a rod of polycrystalline material obtained from the purification process. A rod of an appropriate diameter is held at the top and placed in the crystal-growing chamber. A single crystal seed is clamped in contact at the other end of the rod. The rod and the seed are enclosed in a

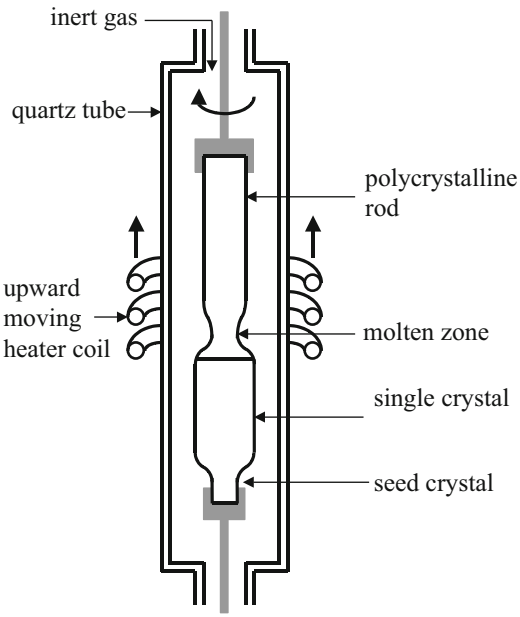


Fig. 17.6 Cross section of a furnace used for the growth of single crystal semiconductor boules by the float-zone process

vacuum chamber or inert atmosphere, and an induction-heating coil is placed around the rod. The coil melts a small length of the rod, starting with part of the single seed crystal. A "float-zone" of melt is formed between the seed crystal and the polysilicon rod. The molten zone is slowly moved up along the length of the rotating rod by moving the coil upward. It should be noted that no crucible is used in this method, as shown in Fig. 17.6. For this reason, extremely high-purity silicon boules, with carrier concentrations lower than 10^{11} cm^{-3}, have been grown by the float-zone method. In general, this method is not used for compound semiconductor growth.

The molten region that solidifies first remains in contact with the seed crystal and assumes the same crystal structure as the seed. As the molten region is moved along the length of the rod, the polycrystalline rod melts and then solidifies along its entire length, becoming a single crystal rod of silicon in the process. The motion of the heating coil controls the diameter of the crystal. Because of the difficulties in preventing the collapse of the molten region, this method has been limited to small-diameter crystals (less than 76 mm). However, since there is no crucible involved in the FZ method, oxygen contamination that might arise from the quartz (SiO_2) crucible is eliminated. Wafers manufactured by this method find their use in applications requiring low-oxygen content, high resistivity starting material for devices such as power diodes and power transistors.

One disadvantage of the float-zone crystal growth is the difficulty in introducing a uniform concentration of dopants. Currently, four techniques are used: core doping, pill doping, gas doping, and finally neutron doping.

Table 17.6 Distribution coefficients for float-zone growth

Impurity	k
B	0.9
P	0.5
Sb	0.07

Core doping uses a doped polysilicon boule as the starting material and then undoped material can be deposited on top of the doped boules until the desired overall doping concentration is obtained. This process can be repeated several times to increase the uniformity or the dopant distribution and, neglecting the first few melt lengths, the dopant distribution is very good. The final dopant concentration of the rod is given by:

$$C(z) = C_c \left[\frac{r_d}{r_f}\right] \left[1 - (1-k)e^{-kz/l}\right] \tag{17.9}$$

where C_c is the dopant concentration in the core rod, r_d is the radius of the core rod, r_f is the radius of the final boule, l is the length of the floating zone, k is the effective distribution coefficient for the dopant, and z is the distance from the start of the boule. Several common distribution coefficients for float-zone growth are shown in Table 17.6

Gas doping simply uses the injection of gases, such as $AsCl_3$, PH_3, or BCl_3, into the polycrystalline rod as it is being deposited or into the molten ring during float-zone refining.

Pill doping is accomplished by inserting a small pill of dopant into a hole that is bored at the top of the rod. If the dopant has a relatively low segregation coefficient, most of it will diffuse into the rod as the melt passes over the rod. Gallium and indium are commonly used as pill dopants.

Finally, light *n*-type doping of silicon can be achieved with neutron bombardment. This is possible because approximately 3.1% of silicon mass is the mass 30 isotope.

17.4.4 Lely Growth Method

Although they account for nearly all bulk semiconductor boules grown commercially, the previously described techniques all make use of the crystallization process from a melt. This is not possible for a number of semiconductor materials, such as silicon carbide (SiC) and the gallium nitride family (GaN and AlN), because they do not have a stable liquid phase under reasonable thermodynamic conditions. SiC melts can only exist under pressures higher than 105 atmospheres and temperatures higher than 3200 °C. Furthermore, under these conditions, the stoichiometry and the stability of the melt could no longer be ensured. At this time, two techniques are being used for the growth of bulk SiC semiconductor boules: the Lely method and

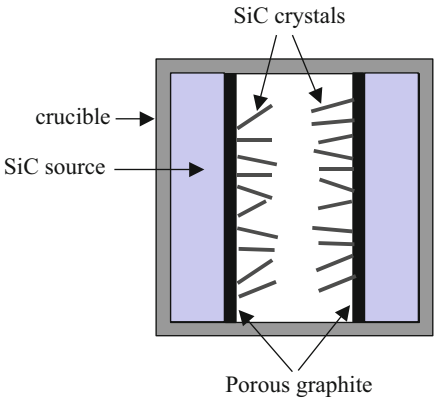


Fig. 17.7 Schematic cross-sectional diagram of a cylindrical crucible used for the Lely growth of SiC

the Modified Lely method. GaN and AlN substrates are usually grown via a hydride vapor-phase epitaxy (HVPE) process.

The Lely growth method is carried out in a cylindrical crucible, schematically depicted in Fig. 17.7. The growth process is basically driven by a temperature gradient which is maintained between the outer and the inner areas of the crucible, with a lower temperature at the center. At the same time, the system is kept under near chemical equilibrium, with lower partial pressures of SiC precursors in the inner colder region. The two areas are separated by porous graphite, which also provides nucleation centers.

The chemical gradient results in a mass transport originating from the outer area toward the inner region. Because the inner region is also colder, SiC will nucleate on the graphite, and crystals will start to grow under their most energetically stable form. Although of the highest quality in terms of possessing low-defect densities, the size of the resulting crystals are somewhat limited and not particularly controllable (typically smaller than 1 cm^2). These crystals are nevertheless used as seed crystals for the Modified Lely method.

The Modified Lely method is the historical name for the Seeded Sublimation Growth or Physical Vapor Transport technique. Its principle is similar to the Lely method with the exception that a SiC seed crystal is used to obtain a controlled nucleation. This method is currently used for the growth of all commercial SiC single crystal boules. A modern crucible for the Modified Lely technique is schematically depicted in Fig. 17.8. The cooler seed is placed on the top to avoid falling contaminants. A polycrystalline SiC source is heated up (up to 2600 °C) at the bottom of the crucible and sublimes at low pressure. Mass transport occurs spontaneously and SiC recrystallizes naturally through supersaturation at the seed.

Although the Modified Lely method is more than 20 years old and has been able to advance the growth of bulk SiC semiconductor crystals, there remain major issues in its process. For instance, the polytype formation and the growth shape are poorly controlled, the doping is nonuniform, and the resulting crystals still have high density of defects, such as micropipes and dislocations.

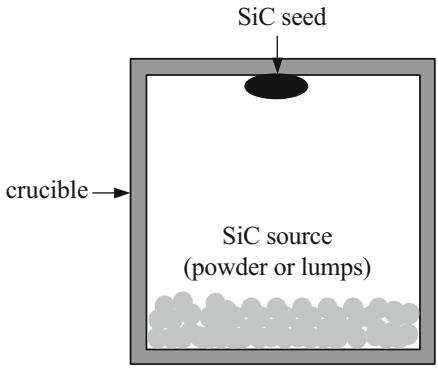


Fig. 17.8 Schematic cross-sectional diagram of a cylindrical crucible used for the Lely growth of SiC

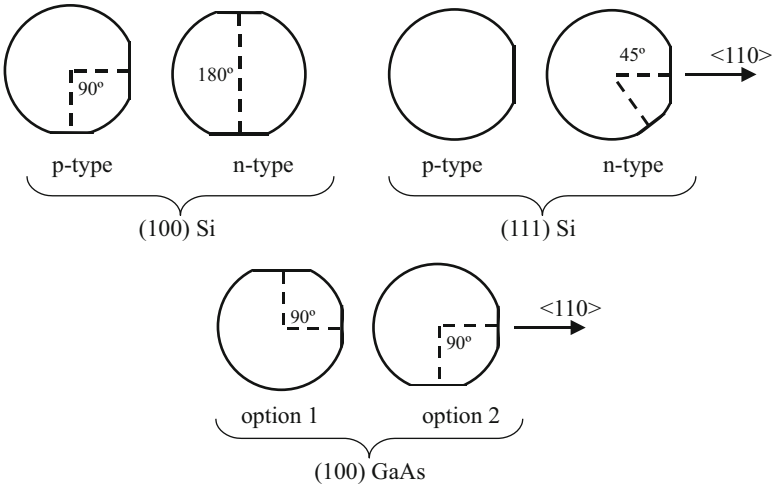


Fig. 17.9 Standard flat orientations for various types of semiconductor wafers. The longer flat is called the primary flat, whereas the shorter one is referred to as the secondary flat

17.4.5 Crystal Wafer Fabrication

After the boule is grown, wafers must be made. Each boule is first characterized for its crystal orientation, dislocation density, and resistivity. Then the seed and the tail of the boule are removed, and the boule is trimmed to the proper diameter. Flats are ground along the entire length of the boule to denote crystal orientation so that the device array can be aligned with respect to the scribe and break directions of the wafer. By convention, the largest or primary flat is ground perpendicular to the ⟨110⟩ direction. Figure 17.9 shows some flat orientations for various types of semiconductor wafers. After grinding the flats, the boule is dipped into an etchant to remove the damage caused by the grinding process. In the last stage, the semiconductor boule is

sliced into wafers using specialized steel or diamond saws. The wafers are then polished to a flat mirror-like surface, chemically etched and cleaned to an atomic cleanliness. All these steps are performed in a clean room with high-purity products in order to avoid any contamination of the surface. Finally, each wafer is individually packaged and sealed in a plastic bag under an inert atmosphere. It is upon such an "epi-ready" (i.e., ready for epitaxy) single crystal that the series of layers needed for a laser or other electronic devices will be deposited.

17.5 Epitaxial Growth Techniques

An overwhelming majority of semiconductor devices, including transistors or diode lasers, require the deposition of a series of thin layers on top of one of the polished wafer substrates previously described. This process of extending the crystal structure of the underlying substrate material into the grown layer is called epitaxy. The term "epitaxy" is a combination of two Greek words, "epi" (placed or resting on) and "taxis" (arrangement or order), and refers to the formation of a single crystal film on top a crystalline substrate. Epitaxy can be further qualified as a function of the nature of the film and the substrate: homoepitaxy is employed when the film and the substrate are made of the same material, and heteroepitaxy is used when the film and the substrate are made of different materials. Homoepitaxy results in a film which is totally lattice matched to the substrate, while heteroepitaxy generally results in a strained or relaxed film depending on the difference of lattice parameters and thermal expansion coefficients between the film and the substrate. An example of homoepitaxy is the growth of a thick GaAs layer (called a buffer layer) on a GaAs substrate in order to improve the quality and purity of the surface prior to the growth of the structure of interest. Examples of heteroepitaxy are the deposition of $In_{0.47}Ga_{0.53}As$ on top of InP substrates (lattice-matched growth) and the growth of GaN on sapphire substrates (lattice-mismatched growth).

The discovery of quantum wells and superlattices has revolutionized the area of semiconductor technology in terms of new devices. These devices require precise control and uniformity of thickness, excellent homogeneity, high purity, very sharp interfaces between the substrate and epitaxial layers, and low misfit dislocations in the epilayers. In the past few decades, epitaxial techniques have advanced to a level where such requirements can be met by a variety of growth methods. These growth techniques include liquid-phase epitaxy (LPE), vapor-phase epitaxy (VPE), metalorganic chemical vapor deposition (MOCVD), and molecular beam epitaxy (MBE), which will be reviewed in the following subsections.

17.5.1 Liquid-Phase Epitaxy

The LPE growth technique uses a system shown in Fig. 17.10 and involves the precipitation of material from a supercooled solution onto an underlying substrate. The LPE reactor includes a horizontal furnace system and a sliding graphite boat.

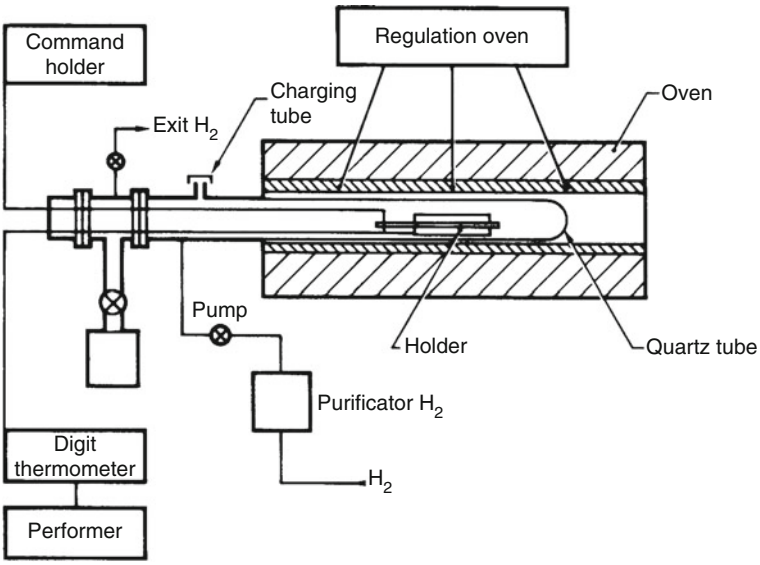


Fig. 17.10 Cross section of a liquid-phase epitaxy system. Inside the horizontal furnace, there is a sliding graphite boat upon which a substrate is held (Copyright 1989 from The MOCVD challenge, vol 1:a survey of GaInAsP-InP for photonic and electronic applications, Razeghi M. p. 6, Fig. 1.2. Reproduced with permission of Routledge/Taylor & Francis Group, LLC)

The apparatus is quite simple, and excellent quality layers with high purity levels can be achieved.

Liquid-phase epitaxy is a thermodynamic equilibrium growth process. The composition of the layers that are grown on the substrate depends mainly on the equilibrium phase diagram and to a lesser extent on the orientation of the substrate. The three parameters that can affect the growth are the melt composition, the growth temperature, and the growth time.

The advantages of LPE are the simplicity of the equipment used, high deposition rates, and the high purity that can be obtained. Background elemental impurities are eliminated by using high-purity metals and the inherent purification process that occurs during the liquid-to-solid phase transition. The disadvantages of the LPE includes a poor thickness uniformity, high surface roughness, melt back effects, and the high growth rates which prevent the growth of multilayer structures with abrupt interfaces. Growing films as thin as a few atomic layers is therefore out of the question using liquid-phase epitaxy, and is usually done using other techniques such as molecular beam epitaxy.

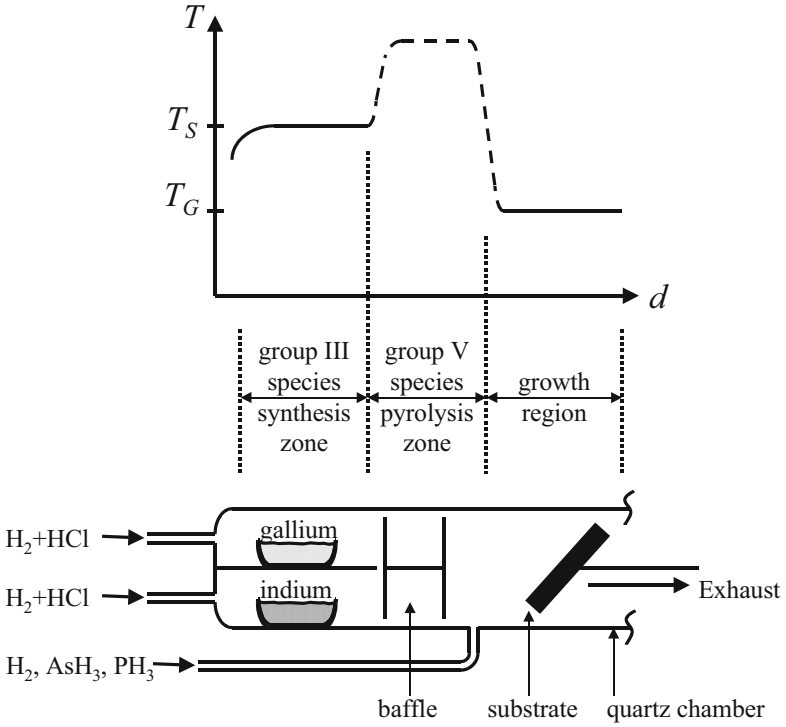


Fig. 17.11 Cross-sectional schematics of a typical VPE reactor, showing the group III species synthesis, group V species pyrolysis, and the growth zones with their respective temperature profiles for the growth of a few selected semiconductors

17.5.2 Vapor-Phase Epitaxy

Like LPE, vapor-phase epitaxy is also a thermodynamic equilibrium growth process. However, unlike LPE, the VPE growth technique involves reactive compounds in their gaseous form. A VPE reactor typically consists of a quartz chamber composed of several zones set at different temperatures using a multi-element furnace, as illustrated in Fig. 17.11.

The group III source materials consist of pure metal elements, such as gallium (Ga) and indium (In), contained in a small vessel. In the first zone, called the group III species synthesis zone, which is maintained at a temperature T_S (~750–850 °C for GaAs or InP growth), the metal is in the liquid phase and reacts with the incoming flow of hydrogen chloride gas (HCl) in the following manner to form group III-chloride vapor compounds which can be transported to the growth region:

$$\mathrm{Ga_{liq}} + \mathrm{HCl_g} \rightarrow \mathrm{GaCl_g} + \frac{1}{2}\mathrm{H_{2_g}}$$

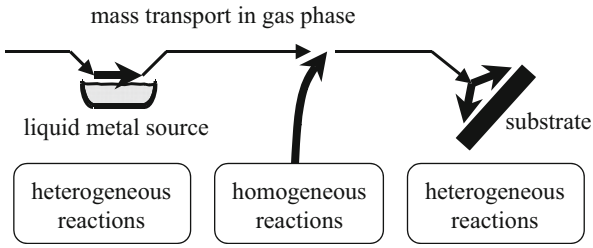


Fig. 17.12 Location of heterogeneous and homogeneous chemical reactions taking place during the vapor-phase epitaxy growth process

$$\mathrm{In_{liq}} + \mathrm{HCl_g} \rightarrow \mathrm{InCl_g} + \frac{1}{2}\mathrm{H_{2_g}}$$

The group V source materials are provided in the form of hydride gases, for example, arsine (AsH_3) and phosphine (PH_3). In the second zone, also called the group V species pyrolysis zone, which is maintained at a temperature $T > T_S$, these hydrides are decomposed into their elemental group V constituents, yielding reactions like:

$$\mathrm{AsH_3} \rightarrow \frac{u}{4}\mathrm{As_4} + \frac{1-u}{2}\mathrm{As_2} + \frac{3}{2}\mathrm{H_2}$$

$$\mathrm{PH_3} \rightarrow \frac{v}{4}\mathrm{P_4} + \frac{1-v}{2}\mathrm{P_2} + \frac{3}{2}\mathrm{H_2}$$

where u and v represent the mole fraction of AsH_3 or PH_3 which is decomposed into As_4 or P_4, respectively.

Finally, in the growth region, which is maintained at a temperature T_G (~680–750 °C for GaAs or InP growth), the group III-chloride and the elemental group V compounds react to form the semiconductor crystal, such as GaAs or InP, onto a substrate.

There are two types of chemical reactions taking place in vapor-phase epitaxy, as illustrated in Fig. 17.12: heterogeneous reactions occur between a solid, liquid, and/or vapor, while homogeneous reactions only occur in the gas phase.

During the growth of a semiconductor film in steady-state conditions, the overall growth process is limited by the heterogeneous reactions. During changes in the composition of the growing semiconductor, for example, when switching the growth from InP to GaInAs, the process is limited by the mass transport in the gas phase.

VPE growth model A simple diffusion model can be developed to gain an understanding of the heterogeneous reactions occurring at the surface of the substrate. Near that surface, there exists a thin stagnant layer, called the boundary layer, which has a thickness δ and within which there is no flow but rather a diffusion of reactants, as shown in Fig. 17.13. The concentration of reactants in the bulk gas phase is denoted C_G, while that at the surface of the substrate is denoted C_S. Two fluxes are considered.

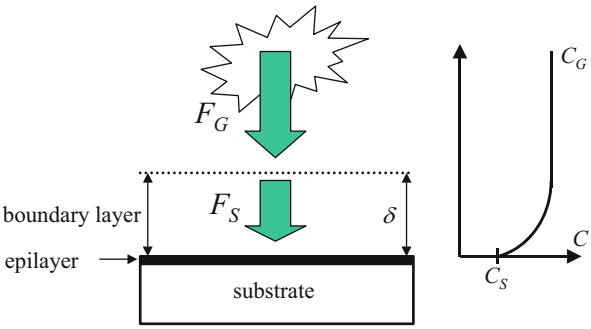


Fig. 17.13 Schematic diagram of the boundary layer near the epilayer/substrate surface in vapor-phase epitaxy. A plot of the concentration of reactants in the bulk gas phase and at the surface as a function of the distance to the substrate is shown on the right

The first one is the flux of molecules from the bulk gas phase onto the sample surface, called F_G. This flux is proportional to the difference between the concentration of reactants C_G and C_S:

$$F_G = \frac{D}{\delta}(C_G - C_S) = h_G(C_G - C_S) \tag{17.10}$$

where D is the effective diffusion coefficient of reactants through the boundary layer, and δ is the distance over which the diffusion is taking place (thickness of the boundary layer). We have also defined a coefficient h_G which is called the vapor-phase mass transfer coefficient.

The second flux, called F_S, corresponds to the incorporation of reactants into the growing crystal. This flux is proportional to the concentration C_S of reactants at the epilayer surface and is given by:

$$F_S = k_S C_S \tag{17.11}$$

where k_S is the surface chemical reaction rate constant. Under steady-state conditions, these fluxes must be equal, i.e., $F_G = F_S$. This translates into the relation between C_S and C_G:

$$C_S = \frac{h_G}{h_G + k_S} C_G \tag{17.12}$$

The growth rate can be calculated as:

$$\frac{dX}{dt} = \frac{F_S}{C} = \frac{h_G k_S}{h_G + k_S} \frac{C_G}{C} \tag{17.13}$$

where we have denoted C the total number of reactants that can be incorporated in a unit volume to form the semiconductor crystal. From this simple expression of the growth rate, we can outline two important growth regimes.

If $h_G \gg k_S$, the growth rate can be approximated by:

$$\frac{dX}{dt} \approx k_S \frac{C_G}{C} \tag{17.14}$$

which means that the surface chemical reaction rate is the limiting step as the growth rate is determined by the surface chemical reaction rate constant k_S.

If $h_G \ll k_S$, the growth rate can be approximated by:

$$\frac{dX}{dt} \approx h_G \frac{C_G}{C} \tag{17.15}$$

which means that the mass transfer is the limiting step as the growth rate is determined by the mass transfer coefficient h_G.

The advantages of VPE include a high degree of flexibility in introducing dopants into the material as well as the control of the composition gradients by accurate control of the gas flows. Localized epitaxy can also be achieved using VPE. One of its main disadvantages is the difficulty to achieve multi-quantum wells or superlattices (periodic heterostructures with a large number of layers having a thickness of the order of a few tens of Angstrom). Other disadvantages include the formation of hillocks and haze, as well as interfacial decomposition during the preheat stage.

17.5.3 Metalorganic Chemical Vapor Deposition

Metalorganic chemical vapor deposition (MOCVD) is a deposition method for the growth of semiconductor thin films. The MOCVD technology has established its ability to produce high-quality epitaxial layers and sharp interfaces and to grow multilayer structures with thicknesses as thin as a few atomic layers.

MOCVD Growth Systems The growth of epitaxial layers from III-V semiconductor compounds is conducted by introducing controlled amounts of volatile compounds of alkyls of group III elements and either alkyls or hydrides of group V elements into a reaction chamber in which a semiconductor substrate is placed on a heated graphite susceptor as depicted in Fig. 17.14. The heated susceptor has a catalytic effect on the decomposition of the gaseous products, such that the semiconductor crystal growth takes place in this hot region.

A typical MOCVD system consists of four major parts: the gas handling system, the reactor chamber, the heating system, and the exhaust and safety apparatus.

The gas handling system includes the alkyl and hydride sources, the valves, pumps, and other instruments necessary to control the gas flows and mixtures. Hydrogen (H_2), nitrogen (N_2), argon (Ar), and helium (He) are the most common inert carrier gases used in the MOCVD growth process. In order to minimize contamination, the gas handling system has to be clean and leak tight. In addition, the material it is made out of must be resistant to the potentially corrosive nature of the sources.

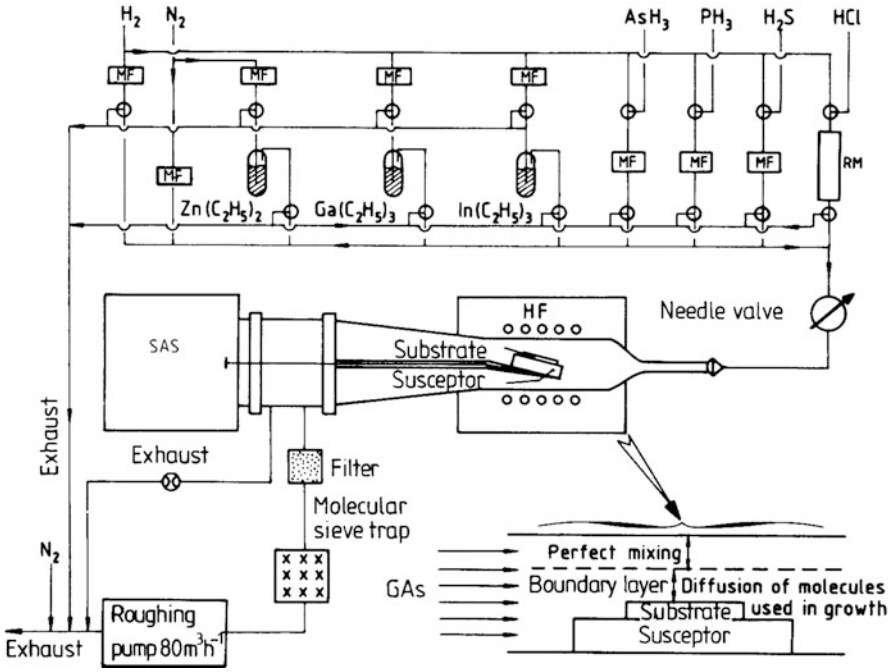


Fig. 17.14 Schematic diagram of a typical low-pressure MOCVD reactor (Copyright 1989 from The MOCVD challenge, vol 1: a survey of GaInAsP-InP for photonic and electronic applications, Razeghi M. p. 13, Fig. 1.7. Reproduced with permission of Routledge/Taylor & Francis Group, LLC)

The purity of the sources is one of the most important issues in modern semiconductor technology. Much effort is constantly devoted to purify every source material used in order to avoid any kind of contamination. Gas purifiers are often used to further purify hydride sources and carrier gases.

Alkyls sources are metalorganic or organometallic compounds, and they are liquids or finely crushed solids usually contained in a stainless steel cylinder called bubbler. The partial pressure of the source is regulated by precisely controlling the temperature and the total pressure inside the bubbler. Electronic mass flow controllers are used to accurately and reliably measure and/or control the mass flow rate of hydride and carrier gases through the gas handling system. Thus, by sending a controlled flow of carrier gas through the bubbler, a controlled mass flow in the form of dilute vapors of the metalorganic compounds can be achieved.

The mixing of volatile compounds in the gas handling system is done in a manifold which first stabilizes the flows and then mixes them and directs them either to the reaction chamber or into the vent (waste). This manifold is designed to uniformly mix metalorganic and hydride source materials prior to reaching the growth zone.

Inside the reaction chamber, the susceptor can be heated using any of the following three methods: radio-frequency (RF) induction heating, radiative (lamp) heating, and resistance heating. The temperature of the substrate is measured using a thermocouple (chromel-alumel) and/or a pyrometer.

The exhaust system may include scrubbing systems, particulate filters, and burnboxes and is aimed at physically or chemically treating the unreacted gases and by-products from the reaction chamber which may still be toxic, pyrophoric, or flammable.

The safety apparatus associated with semiconductor growth systems generally consists of toxic gas monitors used to quantitatively detect the presence of toxic gases such as arsine and phosphine or flammable gases such as hydrogen.

MOCVD Source Materials

A list of suitable metalorganic source materials commonly used in MOCVD, along with their acronyms, some of their physical properties, and their associated safety precautions are listed in Appendix A.10. Examples of suitable hydride precursors for group V, IV, and VI elements, used either to grow the III-V host lattices or to dope the crystals *n*- or *p*-type, are listed in Table 17.7.

For a thorough discussion of these source materials, the interested reader is referred to other books (Razeghi 1989).

MOCVD Growth Process

There exist two types of fundamental processes occurring during crystal growth: thermodynamic and kinetic. Thermodynamics determines the driving force for the overall growth process, and kinetics defines the rates at which the various processes occur. Hydrodynamics and mass transport, which take into account the gas velocities and temperature gradients in the vicinity of the hot susceptor, control the rate of transport of material to the growing solid/vapor interface. The rates of the chemical reactions occurring during growth, either homogeneously in the gas phase or heterogeneously at the growing interface, also play a role. Each of these factors will dominate some aspect of the overall growth process. A study of the dependence

Table 17.7 Hydride source materials for the MOCVD growth and doping of III-V semiconductors. Group IV and VI precursors are generally used for the n-type doping of III-V semiconductors

Name of compound	Acronym	Purpose
Ammonia	NH_3	V element
Arsine	AsH_3	V element
Phosphine	PH_3	V element
Silane	SiH_4	IV element
Disilane	Si_2H_6	IV element
Hydrogen selenide	H_2Se	VI element
Hydrogen sulfide	H_2S	VI element

of a macroscopic quantity, such as growth rate, on external parameters, such as substrate temperature and input precursor (source) flow rates, gives insight into the overall growth mechanism.

Thermodynamic calculations are useful in obtaining information about the solid composition of a multi-component system when vapor-phase compositions are known. They are also useful in obtaining the phase diagram of a multi-component system by calculating the compositions of the crystal for different temperatures and pressures. However, the MOCVD process is by definition not an equilibrium process. Thermodynamics can thus only define certain limits for the MOCVD growth process and is unable to provide any information about the time required to attain equilibrium, the actual steps involved in the pursuit of the lowest-energy state, or the rates of the various processes occurring during the transition from the initial input gases to the final semiconductor solid. These problems can only be approached in terms of kinetics (Stringfellow 1989a).

A much-simplified description of the MOCVD growth process for III-V compounds, such as the growth of GaAs by TMGa and AsH_3, occurring near and at the substrate surface is illustrated in Fig. 17.15. In the first step, both AsH_3 and Ga $(CH_3)_3$ are carried by diffusion through the boundary layer to reach the substrate. The second step involves the surface reactions. The third step is the formation of GaAs, and the final step is the removal of the reaction products.

The growth rate is an important parameter that can be determined from thermodynamic calculations. But, in the MOCVD growth process, the actual growth rate is much lower than that determined from thermodynamics because kinetics and hydrodynamic transport also play a role in determining the growth rate. This is illustrated in Fig. 17.15 which shows the typical growth rate profile as a function of temperature.

For a given flow of source materials, three regimes can be observed for the growth rate. At low temperatures (Fig. 17.16a), chemical reactions at the solid/vapor interface limit the growth rate as they follow an Arrhenius relation of the form exp

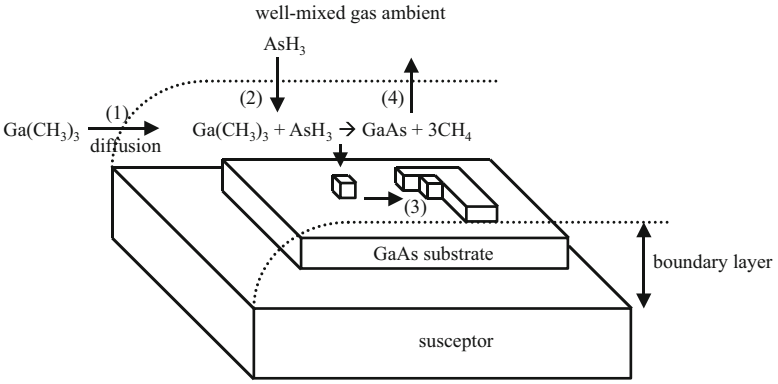


Fig. 17.15 A simplified schematic illustration of the GaAs growth process involving different steps

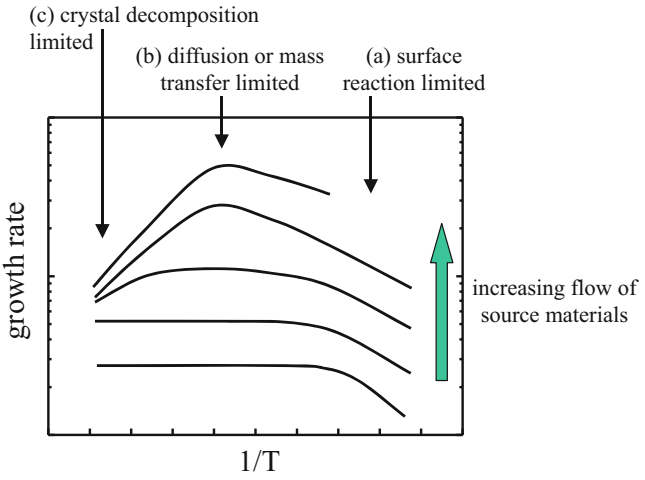


Fig. 17.16 Typical growth rate profile as a function of temperature

$\left(-\frac{E_A}{k_b T}\right)$ where E_A is an activation energy which characterizes the chemical reactions and is of the order of a few eV. For intermediate temperatures (Fig. 17.16b), the growth rate is nearly constant over a wide temperature range. This corresponds to a regime where the diffusion or mass transfer across the boundary layer limits the growth rate. The growth rate is then directly proportional to the flow or partial pressure of incoming source materials and to their diffusion coefficients. In order to achieve a good growth rate control and minimize the sensitivity to temperature, it is preferred to be in conditions which yield a diffusion-limited regime. When the partial pressure of the source materials is increased, the temperature window over which the growth rate is constant is reduced. At high temperatures (Fig. 17.16c), the growth rate becomes independent of temperature and flow of source materials. In this regime, the rate is limited by the decomposition of the growing crystal.

In Situ Characterization Techniques Although the MOCVD growth technique cannot accommodate as many in situ characterization techniques as molecular beam epitaxy (see Subsect. 17.5.4), recent advances in the design and manufacture of MOCVD growth equipment have led to a few viable techniques. Nearly all of them use a laser beam to probe the surface of the growing wafer. One of the pioneering works in this area was done in the late 1980s and consisted of conducting reflectance difference spectroscopy measurements during epitaxial growth (Razeghi 1995). Nowadays, by using a laser with a photon energy lower than the bandgap energy of the growing semiconductor and measuring the intensity of the laser beam reflection, it is possible to qualitatively assess the surface condition, as well as determine the instantaneous thickness of the growing layer.

The MOCVD growth technique has proved to be advantageous in producing some of the highest-quality compound semiconductor materials to date and providing a very high degree of control over the process. MOCVD is also one of the major techniques used in industry, since its process can be fully automated and is capable of yielding the high industrial throughput needed. This has in turn led to the realization of an increasingly large number of high-performance devices, both in electronics and optoelectronics. However, MOCVD still suffers from the large quantity and high toxicity of some of the source materials used, such as arsine and phosphine.

17.5.4 Molecular Beam Epitaxy

Molecular beam epitaxy (MBE) (Cho 1985) is an advanced technique for the growth of thin epitaxial layers of semiconductors, metals, or insulators. A photograph of such a system is shown in Fig. 17.17.

In this technique, the precursor sources are either solids which are sublimated or heated above their melting points in effusion cells or gases which are connected through an injector and cracker. The sources are evaporated in the form of beams of atoms or molecules at a controlled rate onto a crystalline substrate surface held at a suitable temperature under ultrahigh vacuum conditions, as illustrated in Fig. 17.17. The epitaxial layers crystallize through a reaction between the beams originating from the sources and the heated substrate surface. The thickness, composition, and doping level of the epilayer can be very precisely controlled via an accurate control of the beam fluxes. The substrate is mounted on a block and rotated continuously to promote uniform crystal growth on its surface. The beam flux of the source materials is a function of their vapor pressure which can be precisely controlled by their temperature (Fig. 17.18).

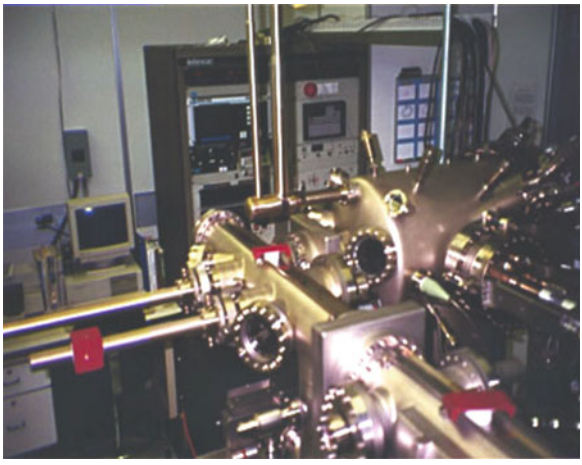


Fig. 17.17 Photograph of a molecular beam epitaxy reactor

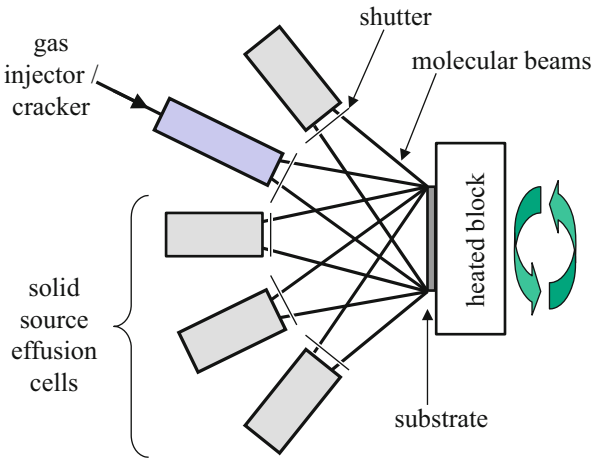


Fig. 17.18 Schematic diagram of an MBE growth system showing a few solid effusion cells, a gas injector/cracker, shutters controlling which sources are used at one time, the path of the beams, and a substrate mounted on a heated block that can be rotated

The thickness, composition, and other properties of the epitaxial layers and heterostructures are directly controlled by the interruption of the unwanted atomic beams with specially designed shutters. An ultrahigh vacuum (UHV) level will ensure the beam nature of the mass flow toward the substrate. This means that the atoms will not interact with each other before reaching the substrate because they have a mean free path longer than the distance between the cells and the substrate. The mean free path Λ of an atom or molecule is expressed as:

$$\Lambda = \frac{1}{\sqrt{2}\pi n d^2} \tag{17.16}$$

in which d is the diameter of the atom or molecule and n is its concentration in the growth chamber given by:

$$n = \frac{P}{k_b T} \tag{17.17}$$

where k_b is the Boltzmann constant and P and T are the pressure and absolute temperature in the MBE growth chamber. The usual distance between the orifice of the cells and the substrate in MBE reactors is about 0.2 m which is two orders of magnitude shorter than the mean free path of atoms or molecules (several tens of meters) at the usual operating pressures in MBE (10^{-5} Pa).

The major difference between MBE and other epitaxial growth techniques stems from the fact that the growth is carried out in an ultrahigh vacuum environment. Therefore, the growth is expected to occur far from thermodynamic equilibrium and is mainly governed by the kinetics of the surface processes. This is in contrast to the other growth techniques, such as liquid-phase epitaxy, where the growth conditions are near the thermodynamic equilibrium and are mostly controlled by diffusion processes near the surface of the substrate. The most important processes in MBE growth occur at the atomic level in the crystallization zone and can be summarized

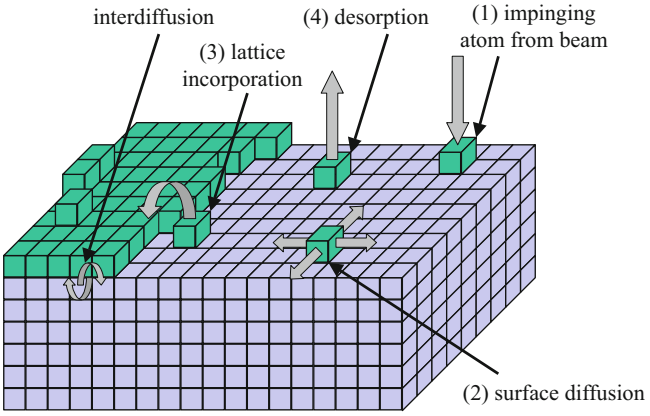


Fig. 17.19 Schematic illustration of the surface processes during MBE epitaxial growth, including (1) the adsorption of the constituent atoms or molecules impinging on the substrate surface, (2) the surface migration and dissociation of the absorbed species, (3) the incorporation of the constituent atoms into the crystal lattice of the substrate or the epilayer, and (4) the thermal desorption of the species not incorporated into the crystal lattice

into four fundamental steps illustrated in Fig. 17.19: (1) adsorption of the constituent atoms or molecules impinging on the substrate surface; (2) surface migration and dissociation of the absorbed species; (3) incorporation of the constituent atoms into the crystal lattice of the substrate or the epilayer, at a site where sufficiently strong bonding exists (that site is usually at the edge of a spreading atomic layer of the growing epitaxial crystal); (4) and thermal desorption of the species that were not incorporated into the crystal lattice.

The atoms and molecules impinging on the substrate are bonded to the surface by weak van der Waals forces and can thus have a high surface mobility when the substrate is adequately heated. However, the growth rate cannot be very high (around one micrometer per hour) because the atoms must be allowed sufficient time to reach their proper position at the step edge before an entire new layer comes down and buries them. Otherwise, we would get a very rough surface with mountain-like and valley-like features on it. Worse yet, the crystal could actually end up with defects, such as missing atoms at sites in the crystal structure that would result in undesirable electrical properties.

Originally, molecular beam epitaxy was a UHV growth technique developed exclusively for solid materials where the cells consisted of a resistively heated crucible in which a piece of solid element was loaded. However, due to the long down time periods necessary to reload the cells and recover the UHV conditions of the system as well as the low growth rates of MBE, some attempts were made to substitute some (if not all) of the solid sources by gas sources that could be changed externally without venting the growth chamber. Nowadays, when all the sources consist of conventional effusion cells containing solid charges of material, the technique is called solid-source MBE (SSMBE). On the other hand, when hydrides

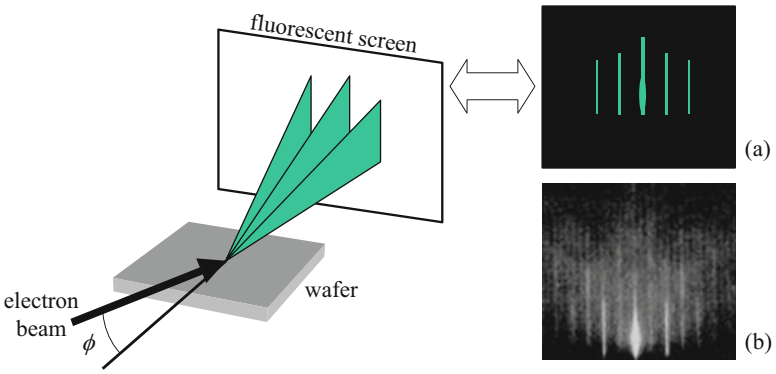


Fig. 17.20 Geometry of RHEED technique. A beam of electrons with an energy in the range 5–50 keV is directed on the substrate surface at an angle ϕ. The electrons are then partially reflected and diffracted by the wafer surface, which leads to the appearance of a bright spot and intensity-modulated streaks on a fluorescent screen, as is schematically shown in (**a**). An actual RHEED pattern is shown in (**b**)

are used instead of solid sources (for group elements, for instance), the name gas-source MBE (GSMBE) is used. When organometallics substitute the solid materials (for group III elements, for instance), the term metalorganic MBE (MOMBE) is employed. But when all the sources are in the gaseous form, the technique is called chemical beam epitaxy (CBE). The main differences between this last technique and MOCVD are the UHV growth conditions and the much smaller quantity of toxic gas which is used during growth, leading to a better acceptance of the technique.

The UHV conditions present in all the MBE techniques also allow the use of in situ diagnostic techniques in order to monitor the growth and substrate surface, such as reflection high-energy electron diffraction (RHEED), Auger electron spectroscopy (AES), X-ray photoelectron spectroscopy (XPS), low-energy electron diffraction (LEED), secondary-ion mass spectroscopy (SIMS), and ellipsometry.

In a RHEED system, a beam of electrons with energies in the range 5–50 keV is directed on the substrate at a grazing angle ϕ (1–2°) as shown in Fig. 17.20. Part of the electrons is directly reflected by the surface, whereas the rest of them are diffracted by the crystalline structure of the epitaxial film. A diffraction pattern, called RHEED pattern, is then formed on a fluorescent screen located on the opposite side of the growth chamber and consists of a bright spot (reflected beam) superposed with intensity-modulated streaks. Since ϕ is very small, the electrons only penetrate into the first atomic layers of the crystal and therefore can only probe a two-dimensional lattice. Therefore, a streaky diffraction pattern is formed instead of the usual spotty pattern which is typical of electron diffraction through a three-dimensional lattice. Since the electrons only penetrate into the first atomic layers of the sample, the RHEED technique is very sensitive to any surface phenomena and can provide useful information about adsorption and desorption of species, roughness, surface reconstruction, substrate miscut, and lattice parameter, in addition to

the general growth parameters such as growth rate and alloy composition. There are two types of RHEED characterization: static and dynamic.

In the first type, the miscut of the substrate, the lattice parameter, and the reconstruction of the surface can be determined from the RHEED diffraction pattern when no growth is occurring. Such information is of particular interest since these parameters directly influence the quality of the growth and also provides useful information about the sample temperature and the strain of the epilayer.

Dynamic RHEED is based on the change of the intensity of the specular beam as a function of the wafer surface roughness, as illustrated in Fig. 17.21. Indeed, during

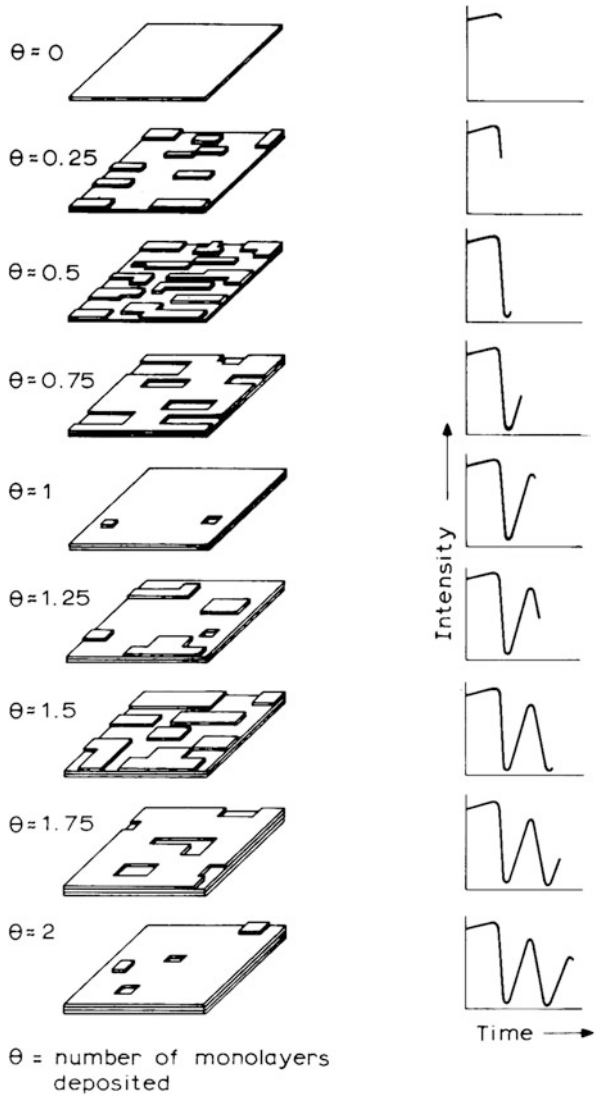


Fig. 17.21 Schematic diagram illustrating the dynamic RHEED process. The sketches on the left show the various stages of the surface morphology during epitaxial growth, while the right plots show the intensity of the RHEED signal from the specular beam as a function of time (Reprinted permission from Elsevier. Surf Sci, vol 168, Joyce BA, Dobson PJ, Neave JH, Woodbridge K, Zhang J, Larsen PK, Bôlger B RHEED studies of heterojunction and quantum well formation during MBE growth-from multiple scattering to band offsets, p. 426, Fig. 2, Copyright 1986)

the epitaxial growth process, starting from an atomically flat surface (i.e., coverage: $\theta = 0$), the roughness increases as a new crystal layer nucleates, thus decreasing the intensity of the reflected beam which is scattered by the increasing number of small islands nucleated on the surface. Once the coverage reaches 50% ($\theta = 0.5$), the roughness is maximal (the intensity of the reflected beam is minimum) after which it will start to decrease as the growing layer is filled, leading to an increase of the intensity of the reflected beam. Once the new layer is completed ($\theta = 1$), the roughness is minimal. The intensity of the specular beam follows this periodic behavior during the growth, with the maximal intensity corresponding to the minimal roughness. The time separation between two adjacent peaks yields the time required for the growth of a single monolayer of the crystal. This is a powerful method which provides an accurate thickness calibration technique that is sensitive to within one single atomic layer.

In spite of its technological advantages over other epitaxial growth techniques, MBE suffers from the high cost to maintain the ultrahigh vacuum environment. In addition, there remain technological challenges, such as increasing the growth rate which remains rather slow and alleviating the difficulty to grow alloys containing phosphorus, such as InP and InGaAsP.

17.5.5 Other Epitaxial Growth Techniques

In general, epitaxial growth is referred to a process in which atoms are randomly deposited on the surface of a substrate and are then properly arranged according to the equilibrium atomic configuration on the surface. Defects are formed when there is a departure from this perfect atomic arrangement. Thus, lateral migration of atoms on the surface aimed at rearranging the surface properly is important in obtaining high-quality epilayers, which is the principle of a special growth technique called migration-enhanced epitaxy (MEE) (Horikoshi 1993).

In the conventional MBE and MOCVD growth of GaAs for instance, Ga and As precursors are introduced onto the substrate surface simultaneously. This leads to the formation of small GaAs islands. In this case, there is an equilibrium density of Ga atoms on the surface. These Ga atoms are very mobile and can migrate on the surface to find more stable sites, before they react with re-evaporated As atoms. However, this process requires high substrate temperatures to guarantee re-evaporation of As atoms. In the absence of As, Ga atoms are even more mobile, and they can migrate even at reduced substrate temperatures. Therefore, high-quality GaAs can be grown after the succeeding As_4 deposition, even at very low substrate temperatures.

Atomic layer epitaxy (ALE) is another peculiar growth technique and is mostly implemented during the MOCVD process (Razeghi 1989). Its main advantage is that it allows for the digital control of the growth rate at a monolayer scale. During an ALE process, the precursors are alternatively injected onto the substrate in the chamber. As a result, gas phase mixing and homogeneous chemical reactions of source materials, commonly found in MOCVD, are suppressed as the growth reaction occurs only on the substrate surface. Therefore, the film thickness can be

controlled with a single atomic layer accuracy. Furthermore, the ALE process exhibits self-limitation, that is, the layer thickness per cycle is independent of subtle variations of growth parameters. The growth rate is only dependent on the number of growth cycles and the lattice constant of the deposited material.

Atomic layer epitaxy is a particular case of self-limiting processes that take place in the gas phase. There exist other types of self-limiting growth processes but using ionic species reactants in solution, in which case the methods are known as successive ionic layer adsorption and reaction (SILAR) or electrochemical ALE (ECALE).

17.5.6 Ex Situ Characterization of Epitaxial Thin Films

Following the epitaxial growth, the semiconductor thin films and structures are removed from the growth system, and their properties are assessed using various ex situ characterization techniques. This is an important quality control step in the development of semiconductor devices, as the quality of the semiconductor material will directly determine the performance of the devices fabricated from it.

Several techniques are commonly employed, such as X-ray diffraction (XRD), scanning and transmission electron microscopy (SEM, TEM), atomic force microscopy (AFM), scanning tunneling microscopy (STM), deep-level transient spectroscopy (DLTS), electrochemical capacitance-voltage measurements (CV), resistivity and Hall measurement, Auger electron spectroscopy (AES), secondary-ion mass spectroscopy (SIMS), photoluminescence (PL), and photoluminescence excitation (PLE). The use of some of them for semiconductor epitaxial thin films has been discussed in detail in Chap. 16.

17.6 Thermodynamics and Kinetics of Growth

In Sect. 17.5.3, thermodynamics and kinetics of MOCVD were briefly introduced. In this section, these two very important topics will be discussed further. Recalling from the MOCVD section, thermodynamics deals with equilibrium conditions and tells us whether or not a chemical reaction is possible. Kinetics, on the other hand, tells us about the rate at which reactions occur. In the following subsections, we will touch upon some of the essential topics involved in the growth of compound semiconductors. These topics include thermodynamics, feasibility of chemical reactions, phase diagrams, and kinetics.

17.6.1 Thermodynamics

In this subsection a brief overview of the thermodynamics of materials will be given. Thermodynamics tells us whether or not a reaction is possible. It can also determine, to some extent, the feasibility of a chemical reaction. In order to get such information, the Gibbs free-energy function, G, is often used:

$$G = H - TS \tag{17.18}$$

where H is the enthalpy, S is the entropy, and T is the absolute temperature. H can be written in terms of the internal energy (E), the volume (V), and the pressure (P) as:

$$H = E + PV \tag{17.19}$$

Now suppose that the initial state of the system (i) changes to a final state (f) due to a chemical reaction, while the temperature is kept constant. The free-energy change can be written as:

$$\Delta G = G_f - G_i = \Delta H - T\Delta S \tag{17.20}$$

The Second Law of thermodynamics states: "In all energy exchanges, if no energy enters or leaves the system, the potential energy of the final state will always be less than that of the initial state ($\Delta G < 0$)." This implies that systems tend to minimize the free energy to a lower value than the initial value. After the system has achieved the equilibrium, ΔG equals 0. For a process that cannot occur, $\Delta G > 0$. Therefore, the possibility of occurrence of a particular reaction can be determined via the sign of ΔG.

17.6.2 Feasibility of Chemical Reactions

For a typical chemical reaction involving materials X, Y, and Z in equilibrium with x, y, and z as the stoichiometric coefficients:

$$xX + yY \rightarrow zZ \tag{17.21}$$

The free-energy change of the reaction is given by:

$$\Delta G = zG_Z - xG_X - yG_Y \tag{17.22}$$

The free energy of individual reactants is often written as:

$$G_i = G_i^0 + RT \ln a_i \tag{17.23}$$

where G_i^0 is the free energy of the species in their standard state and a_i is a term called activity which reflects the change in the free energy when the material is not in its standard state. The standard state is typically 1 atmosphere partial pressure for a gas at 25 °C. A pure liquid or solid is the standard state of the relevant substance. Table 17.8 lists the standard values of the change of enthalpy and entropy for the formation of various substances. Substitution of Eq. (17. 23) into Eq. (17.22) and letting $\Delta G = 0$ yields:

Table 17.8 Standard values of the change of enthalpy and entropy for the formation of some select species at 25 °C and 100 kPa (s = solid, g = gas, l = liquid, aq = aqueous, i.e., dissolved in water)

Species	State	ΔH_f (kJ·mol^{-1})	S (J·mol^{-1}·K^{-1})
H_2O	17.6.2.1.1.1.1.1.1. l	−286	70
H_2O	17.6.2.1.1.1.1.1.2. g	−242	190
CO_2	17.6.2.1.1.1.1.1.3. g	−394	214
O_2	17.6.2.1.1.1.1.1.4. g	0	205
HCl	17.6.2.1.1.1.1.1.5. g	−92	190
HCl	17.6.2.1.1.1.1.1.6. aq	−167	57
H	17.6.2.1.1.1.1.1.7. g	218	115
Cl_2	17.6.2.1.1.1.1.1.8. g	0	220
NaCl	17.6.2.1.1.1.1.1.9. s	−411	72

Table 17.9 Free-energy change and classification of some select reactions

Reactants	Products	ΔG (kJ·mol^{-1})	Classification
$SiH_4 + 2O_2$	$SiO_2 + 2H_2O$	−1307	Highly favorable, highly irreversible
$2SiH_4 + 4NH_3$	$Si_3N_4 + 12H_2$	−742	Favorable, irreversible
SiH_4	$Si + 2H_2$	−57	Moderately favorable, can be reversible
$TiCl_4 + 2NH_3$	$TiN+4HCl + H_2$	+92	Not favorable, possible at elevated temperatures
$TiCl_4 + 2H_2$	Ti + 4HCl	+287	Not favorable, possible only at very high temperatures

$$-\Delta G^0 = RT \ln K \tag{17.24}$$

where

$$K = \frac{a^z_{Z(\text{eq})}}{a^x_{X(\text{eq})} a^y_{Y(\text{eq})}} \tag{17.25}$$

Let us see how thermodynamics can help us find out about feasibility of a chemical reaction. Table 17.8 includes several CVD reactions with different values for the free-energy change term (ΔG). This table shows that oxidation and nitridation of silane are favorable reactions and cannot be reversed, since ΔG is a strongly negative value. Decomposition of silane, however, can be reversible as the reaction has a small value of free-energy change, and, in fact, by adding small amounts of chlorine, the reaction will go the other way. Deposition of TiN is not thermodynamically favorable at room temperature. However, the reaction can take place at slightly higher temperatures (ΔG is a small positive value). As for the deposition of Ti metal, the value of free-energy change is very high. Therefore, much higher temperatures (in excess of 1000 °C) are required for the deposition of Ti.

17.6.3 Phase Diagrams

Phase diagrams allow us to predict and interpret the changes of composition of a material from phase to phase by visual means, i.e., graphs. As a result, phase diagrams have been proven to provide an immense understanding of how a material forms microstructures within itself, leading to an understanding of its chemical and physical properties. Using phase diagrams will allow one to determine which phase or phases are present in a particular system at a given temperature and pressure.

There are a few simple rules associated with phase diagrams with the most important of them being the Gibbs Phase Rule. The Gibbs Phase Rule describes the possible number of degrees of freedom in a (closed) system at equilibrium in terms of the number of separate phases and the number of chemical constituents in the system and can simply be written as:

$$f = C - P + 2 \tag{17.26}$$

where C is the number of components, P is the number of phases, and f is the number of degrees of freedom in the system. The number of degrees of freedom (f) is the number of independent intensive variables (i.e., those that are independent of the quantity of material present) that need to be specified in value to fully determine the state of the system. Typically such variables might be temperature, pressure, or concentration. This rule states that, for a one-component one-phase system, there are two degrees of freedom. For example, on a P–T diagram, pressure and temperature can be chosen independently. On the other hand, for a two-phase system, there is only one degree of freedom, and there is only one pressure possible for each temperature. Finally, for a three-phase system, there exists only one point with fixed pressure and temperature (Fig. 17.22).

17.6.4 Kinetics

As mentioned earlier, thermodynamics deals with the equilibrium processes. It is only concerned with the free energy of the system at its initial and final stages. Only certain limits of the growth process can be defined using thermodynamics: the driving force, maximum growth rate, and the number and compositions of the equilibrium phases. In order to obtain other useful information such as the real growth rate, the actual steps in search of the lowest energy state, or the rate at which various processes occur during the transition from the initial atomic or molecular species to the final solid form, kinetics needs to be considered.

The rate of chemical reactions is usually treated using the theory of absolute reaction rates (Eyring et al. 1941). This theory suggests that, in any chemical reaction, the reactants proceed to products through the formation of an activated complex. For exothermic reactions, the products will have a lower energy than the reactants (Fig. 17.22). The rates of the forward and reverse reactions can be described as:

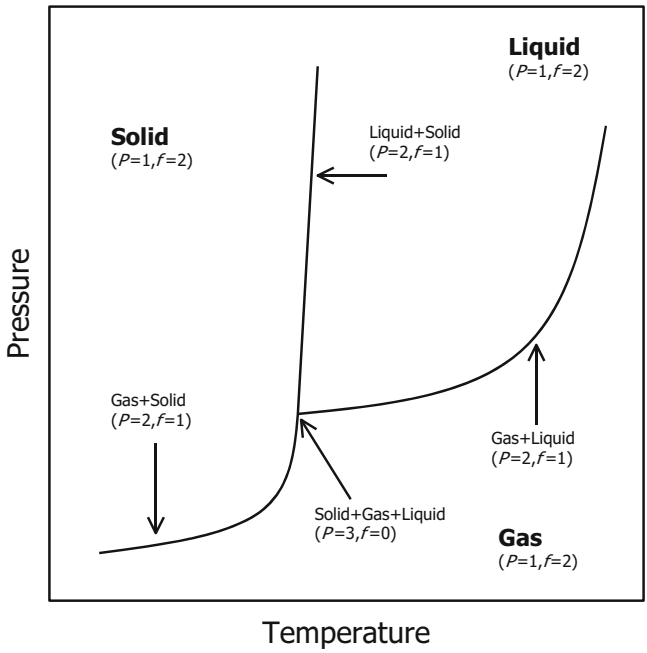


Fig. 17.22 *P–T* diagram of a one-component system showing degrees of freedom for different number of phases

$$\text{Rate} = nk \tag{17.27}$$

where n is the concentration of reactants/products and k is the rate constant usually expressed in terms of the Arrhenius equation:

$$k = Ae^{-E^*/RT} \tag{17.28}$$

In this equation, A is a pre-exponential factor, and E^* is the activation energy of the process. R is the gas constant.

From Fig. 17.23, we find the thermodynamic enthalpy difference from the initial to the final state, ΔH, to be:

$$\Delta H = E_1^* - E_{-1}^* \tag{17.29}$$

At equilibrium, the rates of the forward and reverse reactions are equal:

$$n_i k_1 = n_f k_{-1} \tag{17.30}$$

where subscripts i and f denote the initial and the final state, respectively. The ratio of the concentrations in the final and initial states can be expressed as:

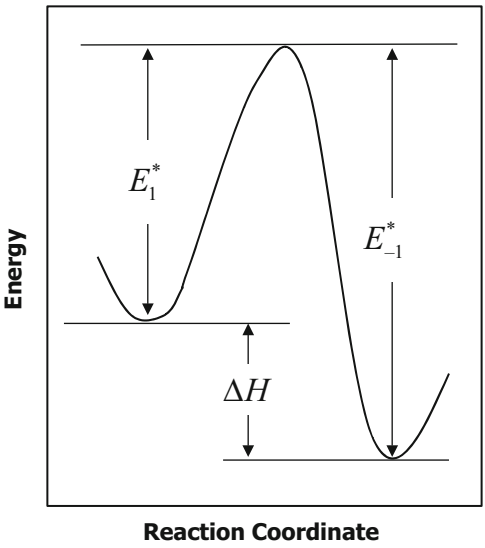


Fig. 17.23 Schematic diagram of energy vs. reaction coordinate. E_1^* and E_{-1}^* are the activation energies of the forward and reverse reactions, respectively

$$\frac{n_f}{n_i} = \frac{k_1}{k_{-1}} = K_1 = \exp\left(\frac{-\Delta G_1^0}{RT}\right) \tag{17.31}$$

In Eq. (17.31), K_1 is the equilibrium constant and ΔG_1^0 is the standard Gibbs free-energy change for the chemical reaction. The standard free-energy change is basically the free energy term (ΔG) under *standard conditions*, which includes a pressure of 1 atmosphere, a temperature of25 °C (298 K), and reactants and products at concentration of 1 mole.

17.7 Growth Modes

Usually growth modes are classified into three categories: the layer-by-layer or Frank-van der Merwe growth mode, the island or Volmer-Weber growth mode, and the layer-plus-island or Stranski-Krastanov growth mode. In lattice-matched systems, the growth mode is determined by the relation between the energies of two surfaces and the interface energy. If the sum of the surface energy (γ_f) of the epitaxial layer and the energy of the interface (γ_i) is lower than the substrate surface energy (γ_s), i.e., $\gamma_f + \gamma_i < \gamma_s$, upon deposition the top material will wet the substrate, leading to the Frank-van der Merwe growth mode (Fig. 17.24a). In other words, in a layer-by-layer growth mode, the deposited atoms are more strongly attracted to the substrate than they are to one another. Most epitaxial techniques take advantage of the Frank-van der Merwe growth mode. Changing the value of $\gamma_f + \gamma_i$ may result in a transition from this growth mode to the Volmer-Weber growth mode where 3D islands are formed (Fig. 17.24b). In this growth mode, the deposited atoms are more

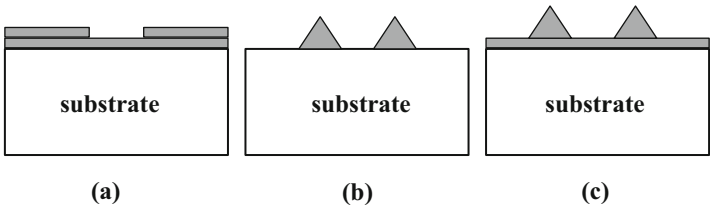


Fig. 17.24 Schematic presentation of the (**a**) Frank-van der Merwe, (**b**) Volmer-Weber, and (**c**) Stranski-Krastanov growth modes

strongly bound to each other than they are to the substrate. A typical example is when a metal is deposited on top of a semiconductor.

In a lattice-mismatched material system, such as GaAs/InAs heterostructures with 7% lattice mismatch, only the first few deposited monolayers form strained epitaxial layers with the lateral lattice constant equal to that of the substrate. When a critical thickness is exceeded, the significant strain occurring in the top layers leads to the spontaneous formation of randomly distributed islands which contribute to relax the elastic energy stored in the system. The phase transition from the two-dimensional epitaxial structure to the random arrangement of three-dimensional islands is called the Stranski-Krastanov transition (Fig. 17.24c). This growth mode is a combination of the other two growth modes and is widely used nowadays to obtain self-assembled quantum dots in lattice-mismatched systems that provide a three-dimensional confinement potential for the carriers.

17.8 Summary

In this chapter, we first reviewed the properties of modern major III-V and II-VI compound semiconductors. By uniformly mixing the various group III and group V elements in the crystal lattice, the lattice parameter and the bandgap energy of the resulting ternary and quaternary alloys can be controlled over a wide range. This is a fundamental property when designing heterostructure compound semiconductor devices. Bulk crystal growth techniques used to synthesize single crystals for today's semiconductor industry were then described. These included the Czochralski, the Bridgman, the float-zone, and the Lely growth methods. We then briefly reviewed the major modern epitaxial growth techniques, such as liquid-phase epitaxy, vapor-phase epitaxy, metalorganic chemical vapor deposition, and molecular beam epitaxy. The advantages and disadvantages of each one have been discussed. These techniques are employed to synthesize semiconductor thin film structures for use in electronic devices. A short overview of thermodynamics and kinetics was given in Sect. 17.6. Finally, the various growth modes were discussed, covering both lattice-matched and lattice-mismatched systems.

Problems

1. From the expressions of the bandgap energy of ternary alloys given in Table 17.2 and using Vegard's law to calculate their lattice parameters, plot the energy bandgap of the following ternary alloys as a function of their lattice parameter:
   $Al_xGa_{1-x}As$, $Al_xIn_{1-x}As$, $Ga_xIn_{1-x}P$, $Ga_xIn_{1-x}As$, GaP_xAs_{1-x}, and InP_xAs_{1-x}.
2. Derive the relation given in Eq. (17.3):

$$a_{A_xB_{1-x}C_yD_{1-y}} = xya_{AC} + x(1-y)a_{AD} + (1-x)ya_{BC} + (1-x)(1-y)a_{BD}$$

3. (a) What is the relationship between the Al mole fraction (x) and the In mole fraction (y) of quaternary $Al_xIn_yGa_{1-x-y}N$ if it is to be lattice-matched to GaN? The lattice parameter of $Al_xIn_yGa_{1-x-y}N$ is given as:

$$a(Al_xIn_yGa_{1-x-y}N) = (1-x-y)a_{GaN} + xa_{AlN} + ya_{InN}.$$

   The lattice parameters of GaN, AlN, and InN are 3.189, 3.112, and 3.545 Å, respectively.
   (b) Using a similar expression as above to calculate the bandgap energy of the quaternary $Al_xIn_yGa_{1-x-y}N$ in terms of its constituent binary compounds, find the chemical formula of the quaternary material of part (a) if the wavelength of the emitted light is to be 300 nm. The bandgap energies of the binary compounds are given as: $E_g(GaN) = 3.4$ eV,
   $E_g(AlN) = 6.0$ eV, and $E_g(InN) = 0.7$ eV

$$E_g\left(Al_xIn_yGa_{1-x-y}N\right) = (1-x-y)E_g(GaN) + xE_g(AlN) + yE_g(InN)$$

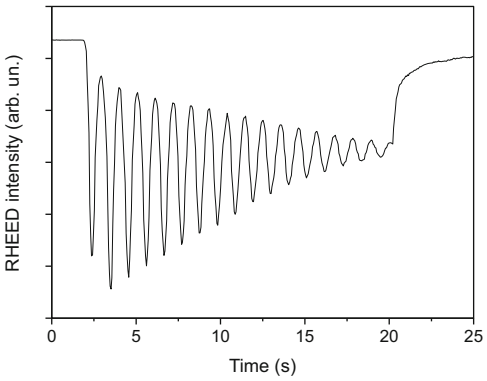
4. Using the diagram in Fig. 17.1, graphically determine the compositions x and y of the quaternary alloy $Ga_xIn_{1-x}P_{1-y}As_y$ which would yield a bandgap energy corresponding to the following wavelengths while being lattice matched to either InP or GaAs: 808 nm, 980 nm, 1.3 μm, 1.55 μm.
5. Compare the MBE and MOCVD growth techniques, using a table that shows some of the advantages and disadvantages of each method.
6. Derive Eq. (17.8): $C_s = kC_0(1-X)^{k-1}$, where:

   C_s = impurity concentration in the solid
   C_0 = original impurity concentration in the melt
   k = segregation coefficient
   X = fraction of the melt that has solidified

7. Plot the dopant concentration profile of a 20″ long silicon rod grown by the float-zone technique using P as a dopant in a core doping scheme for various lengths of the floating zone. Assume the dopant concentration in the core to be 10^{19} cm^{-3} and the radius of the core to be four times smaller than that of the final rod. Which one results in a more uniform doping profile: a long float-zone or a short one?
8. Determine the growth rate of a layer grown by MOCVD using the following parameters:

Diffusion coefficient (D) = 5×10^{-6} cm^2·s^{-1}
Thickness of the boundary layer (d) = 5 mm
Surface reaction chemical rate constant (k_s) = 10^{-3} cm·s^{-1}
Concentration of reactants in gas phase (C_G) = 10^{18} cm^{-3}
Maximum number of reactants incorporating into the crystal (C) = 10^{20} cm^{-3}

9. The figure represents the RHEED oscillation during homoepitaxy of GaAs in an MBE system.
   (a) At what moment did the growth start and stop?
   (b) What is the total thickness of GaAs material deposited?
   (c) Give an estimation of the growth rate, in monolayer per second and in micrometer per hour.



10. (a) Why does the amplitude of the oscillation slowly decrease with time in the figure of last problem?
    (b) Why does the RHEED intensity increase at the end of the curve?
11. In MBE, the deposition of $Al_xGa_{1-x}As$ is performed by opening simultaneously the Ga, Al, and As shutters.
    (a) Since, in normal growth conditions, the incorporation of Al and Ga atoms is unity, find an expression for the Al composition as a function of the growth rate of GaAs, AlAs, and AlGaAs.
    (b) How would you determine the Al fraction with the RHEED system?

References

Casey HC Jr, Panish MB (1978) Heterostructure lasers, parts A & B. Academic Press, New York
Cho AY (1985) The technology and physics of molecular beam epitaxy. Plenum Press, New York
Glasstone S, Leidler KJ, Eyring H (1941) The theory of rate processes. McGraw-Hill, New York
Horikoshi Y (1993) Migration-enhanced epitaxy of GaAs and AlGaAs. Semicond Sci Technol 8:1032–1051
Ilegems M, Panish MB (1974) Phase equilibria in III-V quaternary systems-application to Al-Ga-P-As. J Phys Chem Solids 35:409–420

Jordan AS, Ilegems M (1974) Solid-liquid equilibria for quaternary solid solutions involving compound semiconductors in the regular solution approximation. J Phys Chem Solids 36:329–342
Joyce BA, Dobson PJ, Neave JH, Woodbridge K, Zhang J, Larsen PK, Bôlger B (1986) RHEED studies of heterojunction and quantum well formation during MBE growth-from multiple scattering to band offsets. Surf Sci 168:426
Ray B (1969) II-VI compounds. Pergamon Press, Oxford, UK
Razeghi M (1989) The MOCVD challenge volume 1: a survey of GaInAsP-InP for photonic and electronic applications. Adam Hilger, Bristol
Razeghi M (1995) The MOCVD challenge volume 2: a survey of GaInAsP-GaAs for photonic and electronic device applications. Institute of Physics, Bristol, pp 21–29
Roberts GG, Zallen R (1971) Quenching of photoconductivity and luminescence in natural crystals of mercury sulphide. J Phy Chem: Solid State Phys 4:1890–1897
Stringfellow GB (1989a) Organometallic vapor-phase epitaxy: theory and practice. Academic Press, Boston
Van Vechten JA, Bergstresser TK (1970) Electronic structures of semiconductor alloys. Phys Rev B 1:3351–3358

Further Reading

Gise PE, Blanchard R (1979) Semiconductor and integrated circuit fabrication techniques. Reston Publishing, Reston
Middleman S, Hochberg AK (1993) Process engineering analysis in semiconductor device fabrication. McGraw-Hill, New York, pp 230–257
Nakajima K (1982) Liquid-phase epitaxy. In: Pearsall TP (ed) GaInAsP alloy semiconductor. Wiley, Chichester, pp 43–59
Ohring M (1992) The materials science of thin films. Academic Press, San Diego
Razeghi M (1990) LP-MOCVD growth, characterization, and application of InP material. Semicond Semimetals 31:256–257
Stringfellow GB (1989b) Organometallic vapor-phase Epitaxy: theory and practice. Academic Press, Boston
Tu KN, Mayer JW, Feldman LC (1992) Electronic thin films science for electrical engineers and materials scientists. Macmillan Publishing, New York

Semiconductor Characterization Techniques

18

18.1 Introduction

Semiconductor characterization techniques are used in order to gain knowledge on the physical properties of a semiconductor crystal. The process is similar to decoding the DNA sequence of a living organism as it involves understanding the nanoscale structure of the crystal, i.e., its atoms, electrons, structures, and interactions with the surrounding environment. The knowledge gained from the characterization process is essential in determining whether the semiconductor crystal probed is suitable for a particular device component with certain functionalities.

Semiconductor characterization is generally initiated immediately after the synthesis of a crystal. We can distinguish three types of characterization techniques: structural, optical, and electrical. In this chapter, we will briefly review the most common of these semiconductor characterization techniques. The discussion and examples will be primarily directed toward semiconductor thin films, although most of the same techniques can be readily used for bulk crystals as well.

18.2 Structural Characterization Techniques

18.2.1 X-ray Diffraction

X-ray diffraction employs electromagnetic waves with a wavelength on the order of one angstrom. Since wave diffraction occurs when the dimensions of the diffracting object are of the same order of magnitude as the wavelength of the incident wave, X-rays are ideally suited to probe crystal lattice structures.

X-ray diffraction of semiconductor thin films is generally carried out in a diffractometer. The source of the X-rays is called an X-ray tube (Fig. 18.1) and consists of a water-cooled copper target onto which an accelerated electron beam (up to a few 10's of keV) is impinging inside a vacuum tube. Because of the bremsstrahlung

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_18

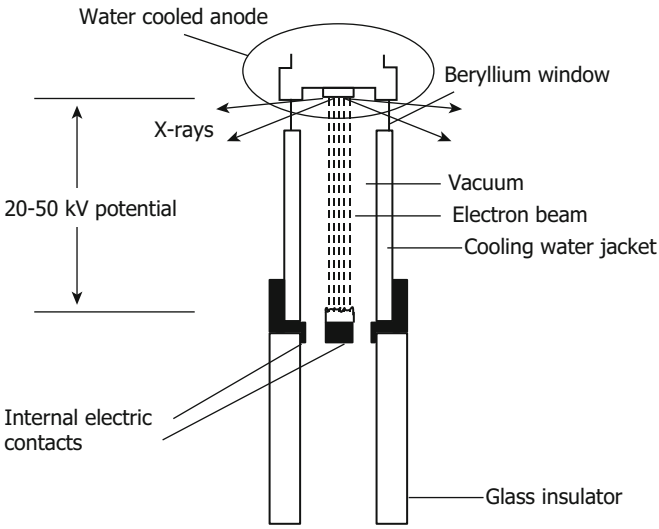


Fig. 18.1 Schematic diagram of an X-ray tube

effect, X-rays are emitted with wavelengths that are characteristic of the copper element. Bremsstrahlung is the original German name for the effect of generation of X-rays via electron deceleration through its interaction with the Coulomb field of the nucleus (of copper, in this case). Through these inelastic interactions, X-rays are emitted which can have energies as high as the beam energy. These X-rays are then filtered and collimated into a beam through the use of a monochromator consisting of nearly perfect silicon crystals placed at specifically chosen angles to permit reflection of the X-rays.

Diffracted waves from different atoms can interfere with each other, and the resultant intensity distribution is strongly modulated by this interaction. If the atoms are arranged in a periodic fashion, as in crystals, the diffracted waves will consist of sharp interference maxima (peaks) with the same symmetry as in the distribution of atoms. Measuring the diffraction pattern therefore allows us to deduce the distribution of atoms in a material.

The peaks in an X-ray diffraction pattern are directly related to the atomic distances. For a given set of lattice planes with an interplane distance d, the condition for a diffraction (peak) to occur can be found using Bragg's law:

$$2d \sin\theta = n\lambda \tag{18.1}$$

where θ is the incident angle, λ is the wavelength of the X-ray, and n is an integer representing the order of the diffraction peak. This process is shown schematically in Fig. 18.2.

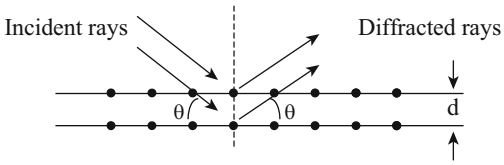


Fig. 18.2 Schematic of diffraction of X-rays by a crystal

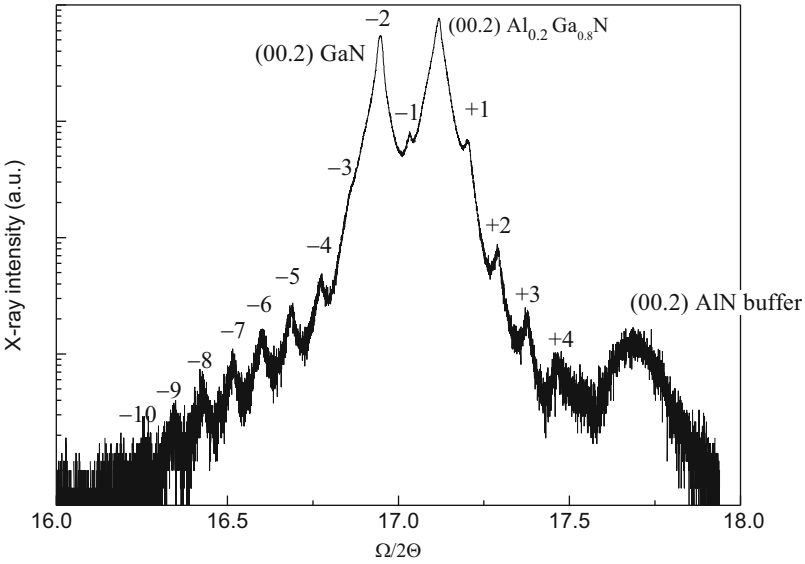


Fig. 18.3 X-ray curve of an $Al_{0.2}Ga_{0.8}N$/GaN superlattice grown on GaN/AlN buffer layer. The individual $Al_{0.2}Ga_{0.8}N$, GaN, and AlN peaks as well as the superlattice satellite peaks are clearly discernible on the graph

Figure 18.3 shows an X-ray diffraction curve of an $Al_{0.2}Ga_{0.8}N$/GaN superlattice structure grown on a GaN template layer. X-ray diffraction measurements on semiconductors can yield useful information such as:

- Lattice constants: The mismatch between the epilayer and the substrate perpendicular to the growth plane can be determined, which is also indicative of strain and stress.
- Rocking curve: The width of the X-ray rocking curve, also called full width at half maximum (FWHM) in units of arcsec or arcmin, is inversely related to the number of dislocations in the epilayer. Therefore this measurement can be used as a measure of the film quality.
- Thickness and quality of superlattices: Thickness of the various layers in multilayer structures like superlattices can be determined by the distance between the satellite peaks appearing on the sides of the main peak. Also the intensity and number of satellite peaks are measures of the film quality.

18.2.2 Electron Microscopy

Scanning Electron Microscopy

A scanning electron microscope (SEM) is probably the most widely used semiconductor characterization instrument. A schematic of a typical SEM system is shown in Fig. 18.4. Electrons are emitted from a tungsten cathode either thermionically or via field emission and are focused by two successive condenser lenses into a very narrow beam. Two pairs of coils deflect the beam over a rectangular area of the specimen surface. Upon impinging on the specimen, the primary electrons transfer their energy inelastically to other atomic electrons and to the lattice. Through many random scattering processes, some electrons manage to leave the surface to be collected by a detector facing the specimen. Usually these are the secondary electrons, originated from a depth of no larger than several angstroms, that are collected by the detector. A photomultiplier tube (PMT) amplifier is used to amplify the signal, and the output serves to modulate the intensity of a cathode ray tube (CRT). Research-quality SEMs are generally able to produce images with a resolution of ~50 Å.

SEM not only can provide images of the surface but also, by rotating the sample, one can obtain information about the thickness of various layers in the structure (cross-sectional SEM). Figure 18.5a illustrates a bird's eye view image of a surface of a "nanopillar" sample, while Fig. 18.5b displays the cross section of a multilayer semiconductor structure.

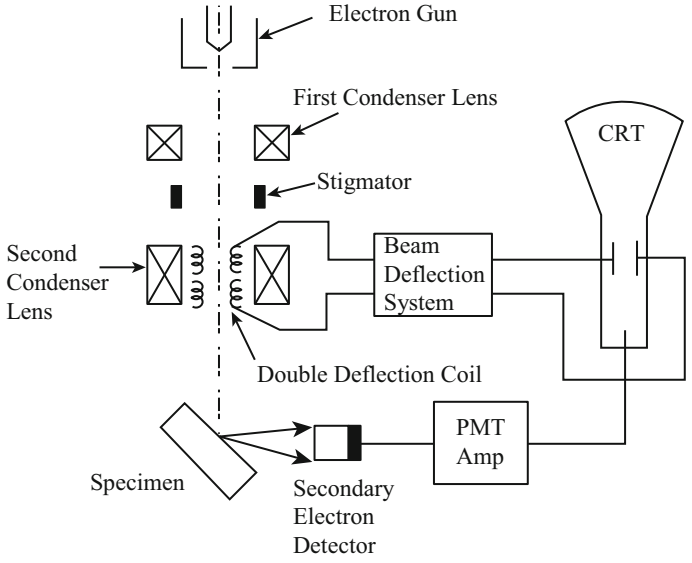


Fig. 18.4 Schematic of a scanning electron microscope

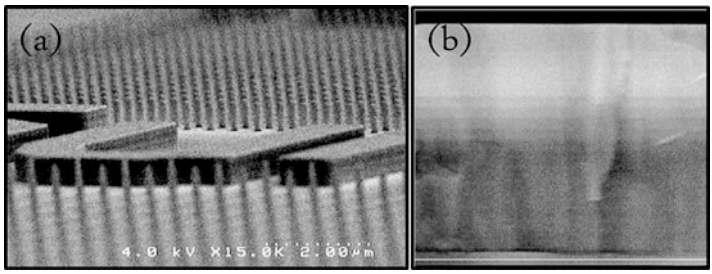


Fig. 18.5 (**a**) Bird's eye view of the surface of a nanopillar sample and (**b**) cross-sectional SEM image of a multilayer semiconductor structure

Transmission Electron Microscopy

Transmission electron microscopy (TEM) is a complex characterization technique that takes advantage of electron diffraction to give the user valuable information regarding the crystallography of the films and, in the image mode, provide high-resolution images of both plain-view and cross-sectional view of the films. A variety of useful information, such as defect structures, structure of grain boundaries, phase identification, crystallographic orientation, quality of the interfaces, etc., can be obtained using this technique.

Figure 18.6 shows the two basic modes of operation of TEM, image mode and diffraction mode. Electrons are thermionically emitted from the gun and are accelerated to high voltages (in excess of 100 keV). A condenser lens section projects the electron beam onto the specimen. Two types of scattering can occur when electrons hit the specimen: elastic scattering results in no loss of energy, while inelastic scattering involves some energy loss. Diffraction patterns can be obtained from elastically scattered electrons, while inelastically scattered electrons give rise to a spatial variation in the intensity of the transmitted beam. Inelastic interactions between the electron beam and the specimen at grain boundaries, dislocations, defect sites, density variations, etc. are the cause of inelastic scattering. Figure 18.5 shows a high-resolution lattice image of the AlN/Al_2O_3 interface. Dislocations can be identified when any of the atomic planes terminates (Fig. 18.7).

TEM is capable of producing high magnifications, due to the small effective wavelengths that are used. Recalling de Broglie's relation from (Eq. 3.3):

$$\lambda = \frac{h}{p} \tag{18.2}$$

As mentioned above, electrons are accelerated to very high energies. If we let this potential energy, eV, equal the kinetic energy of the electrons:

$$\mathrm{eV} = \frac{m_0 v^2}{2} \tag{18.3}$$

the momentum of an electron can be written as:

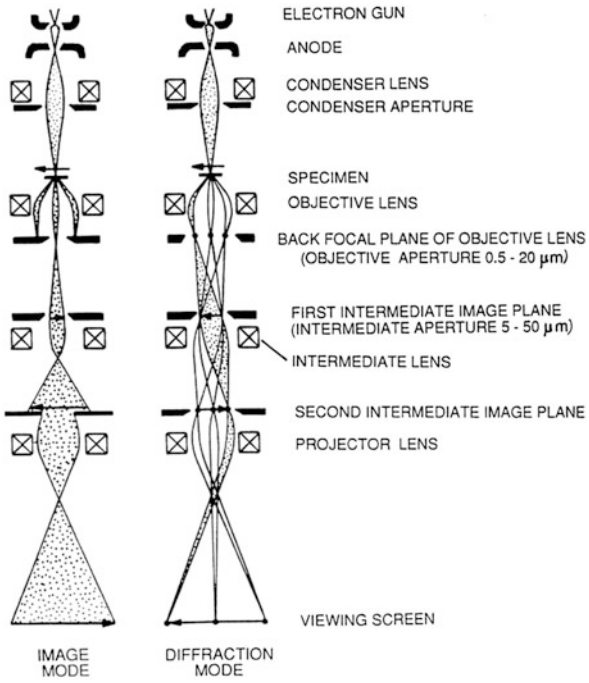


Fig. 18.6 Schematic of the TEM in imaging and diffraction modes. (Reprinted from Thomas G, Goringe MJ Transmission electron microscopy of materials, Fig. 6.10. Copyright 1979 by John Wiley and Sons. Reprinted with permission of CBLS)

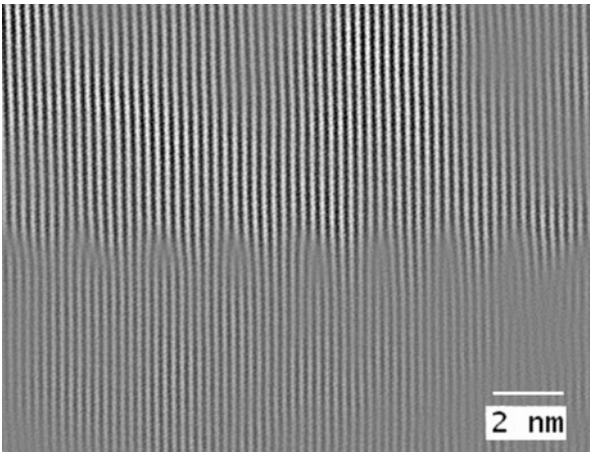


Fig. 18.7 High-resolution TEM image of the interface of AlN and sapphire (Al_2O_3). One misfit dislocation generates when an atomic plane ends

$$p = m_0 v \tag{18.4}$$

Therefore, the wavelength of the electrons, from the above three equations, can be expressed as:

$$\lambda = \frac{h}{\sqrt{2m_0 \mathrm{eV}}} \tag{18.5}$$

For instance, if the acceleration energy of 100 keV is applied, the wavelength will be as small as 0.0386 Å. It should be noted that at such high energies, the velocity of the electrons becomes comparable with the velocity of light. Therefore, in order to have a more accurate evaluation of the wavelength, relativistic effects have to be considered. The modified expression is:

$$\lambda = \frac{h}{\sqrt{2m_0 \mathrm{eV}\left(1 + \frac{\mathrm{eV}}{2m_0 c^2}\right)}} \tag{18.6}$$

For example, with an acceleration voltage of 1 MV, the nonrelativistic wavelength is 0.0122 Å, while the relativistic value is only 0.0087 Å (Williams and Carter 1996).

18.2.3 Energy Dispersive Analysis Using X-rays (EDX)

In EDX an electron from an outer shell of an atom (e.g., the 2*s* shell) lowers its energy to fill the hole in a lower shell (e.g., the 1*s* shell) which results in the emission of an X-ray. These emitted X-rays are characteristic of the particular atom undergoing emission. Therefore, by looking at the X-ray spectral lines of an atom, one could identify that specific atom.

Majority of EDX systems are interfaced to SEM, where they use the same electron beam source to excite X-rays from the specimen under study. A cooled Si (Li) detector (lithium drifted silicon detector) is used to detect X-rays. An emitted X-ray from a specimen generates a photoelectron upon interception by the detector. This photoelectron in turn generates an electron-hole pair. The number of electron-hole pairs, or equivalently the amplitude of the generated voltage pulse, is proportional to the incident photon energy. After amplifying, sorting, counting, and storing the pulses within a range of voltages (energies), the final spectrum will be plotted. Figure 18.8 shows an example of an EDX plot.

18.2.4 Auger Electron Spectroscopy (AES)

The AES technique takes advantage of the Auger transitions that were introduced in Chap. 8. In an Auger process, three electron levels are involved: an electron from an outer level lowers its energy to fill a hole. Instead of generating a photon, this process can result in the ejection of an electron from a third level. The electron that leaves the atom is called the Auger electron. Similar to EDX, the particular atom under test can be identified by looking at the Auger spectral lines.

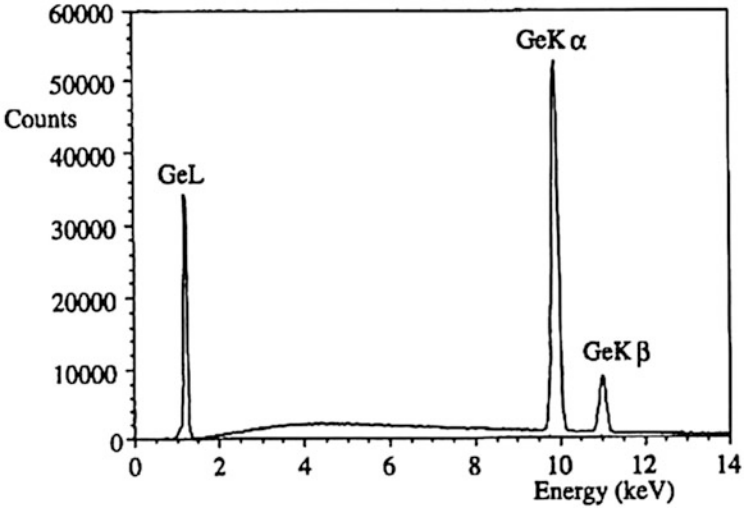


Fig. 18.8 An example of an EDX measurement. Multiple lines of Ge emission correspond to the various electron energy transitions. (Reprinted with permission of Springer Science and Business Media. Williams DB, Carter CB Transmission electron microscopy, p. 557, Fig. 32.2. Copyright 1996 Plenum Press, New York)

A typical Auger spectrometer is kept under ultrahigh vacuum (10^{-10} Torr level) to avoid contaminations. A focused electron beam source of ~2 keV in energy is scanned over the sample area under test. The emitted Auger electrons are then analyzed by an analyzer. The Auger peaks are barely distinguishable above the background signal; therefore, in order to accentuate the energy and magnitude of these peaks, the differentiated signal is generally plotted, as shown in Fig. 18.9.

18.2.5 X-ray Photoelectron Spectroscopy (XPS)

In the XPS technique, low-energy X-rays are used as a source rather than electrons in the case of EDX and AES. Electrons are ejected when the photon is absorbed via the photoelectric effect. In this case the energy of the ejected electron can be written as:

$$E_{\mathrm{KE}} = h\upsilon - E_{\mathrm{BE}} \tag{18.7}$$

where E_{KE} is the energy of the ejected electron, hυ is the energy of the incident photon, and E_{BE} is the energy of the involved bound electron state. By measuring the photoelectron energy, it will be possible to identify the particular atom, since the values of binding energy are element specific. An example of an XPS spectrum (for silver (Ag)) is shown in Fig. 18.10. It should be noted that for multicomponent samples the intensities of the peaks are proportional to the concentration of the element within the sampled region.

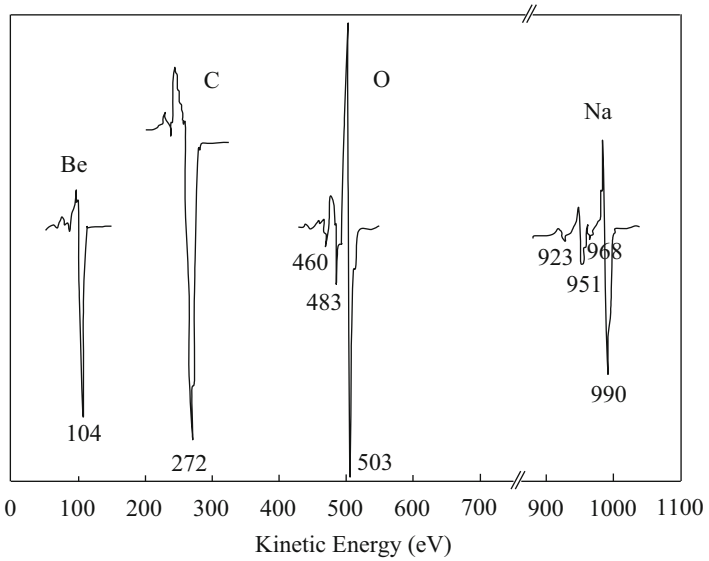


Fig. 18.9 Auger electron spectra of various elements

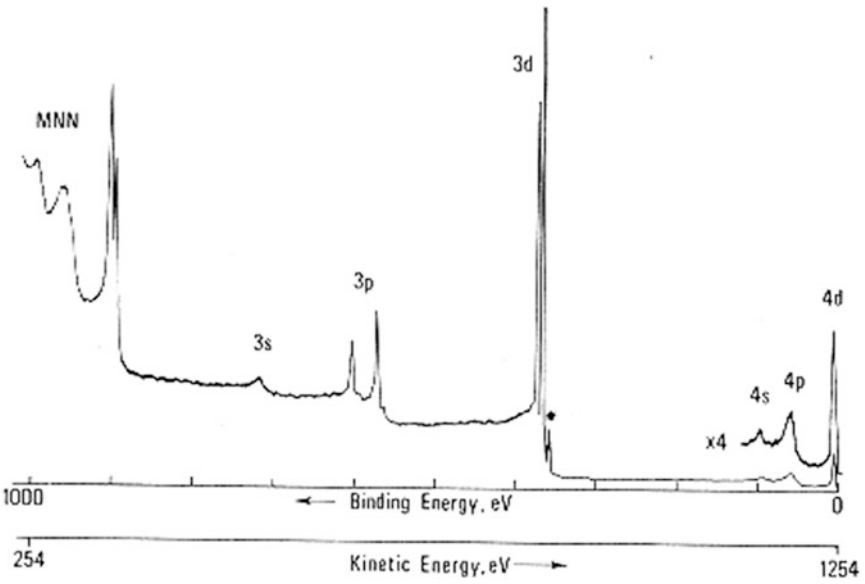


Fig. 18.10 An XPS spectrum from a silver sample (Reproduced with permission. Briggs D, Seah MP Practical surface analysis: by Auger and X-ray photoelectron spectroscopy, p. 112, Fig. 3.16. Copyright 1983 John Wiley & Sons Limited)

18.2.6 Secondary-Ion Mass Spectroscopy (SIMS)

SIMS is a technique used to identify and quantify various types of atoms on the surface or inside a solid sample. In SIMS the material is bombarded by a beam of high-energy ions (1~30 keV) resulting in the ejection or sputtering of atoms from the material. A small percentage of these ejected atoms leave as either positively or negatively charged ions, which are referred to as "secondary ions."

These sputtered secondary ions are then collected and analyzed by a mass-to-charge spectrometer. Elements are identified through their atomic mass values, while their concentration is determined by counting the number of corresponding secondary ions.

The sensitivity of a SIMS measurement is dependent upon the yield of secondary-ion sputtering, which in turn depends on the material under study, the specimen's crystallographic orientation, and the nature, energy, and incidence angle of the primary beam of ions. The proper choice of primary ion beam is therefore important in enhancing the sensitivity of SIMS. O_2^- atoms are usually used for sputtering electropositive elements or those with low ionization potentials such as Na, B, and Al. On the other hand, Cs^+ atoms are better at sputtering negative ions from electronegative elements such as C, O, and As. The detection limit of SIMS is severely reduced with improper selection of the ion beam. Liquid metal ion sources are used for high-resolution measurements, since they can provide smaller beam diameters.

Two types of SIMS are usually considered: "static" SIMS works with low-energy ion sources (0.5–3 keV) which result in low sputter rates (in units of monolayers per second). This mode of operation is suitable for surface analysis, since it will take a long time for the surface to be modified by ion bombardment. "Dynamic" SIMS, on the other hand, uses high-energy ion beams (higher than 3 keV) which results in high sputter rates. This mode of operation is suited for depth profile analysis of the sample under test. Figure 18.11 shows a SIMS depth profile of a GaN sample showing its concentration of impurities (oxygen, carbon, silicon) using Cs^+ bombardment.

18.2.7 Rutherford Backscattering (RBS)

In the RBS technique, very high-energy beams (in the MeV range) of low mass ions (He, C, N, etc.) are accelerated, collimated, and focused upon the sample under test. These high-energy beams have the ability to penetrate deep into the sample (several microns). Such beams cause little sputtering of the surface atoms. Sometimes they penetrate the atomic electron cloud shield and collide with the nuclei of the target atoms. The result is an elastic scattering from the Coulomb repulsion between ion and nucleus, known as Rutherford scattering.

From energy and momentum conservation laws, we know that if an incident ion of mass M_0 and energy E_0 hits a surface atom of mass M, the elastic collision will cause the ion to have an energy E_1 afterward given by Ohring (1992):

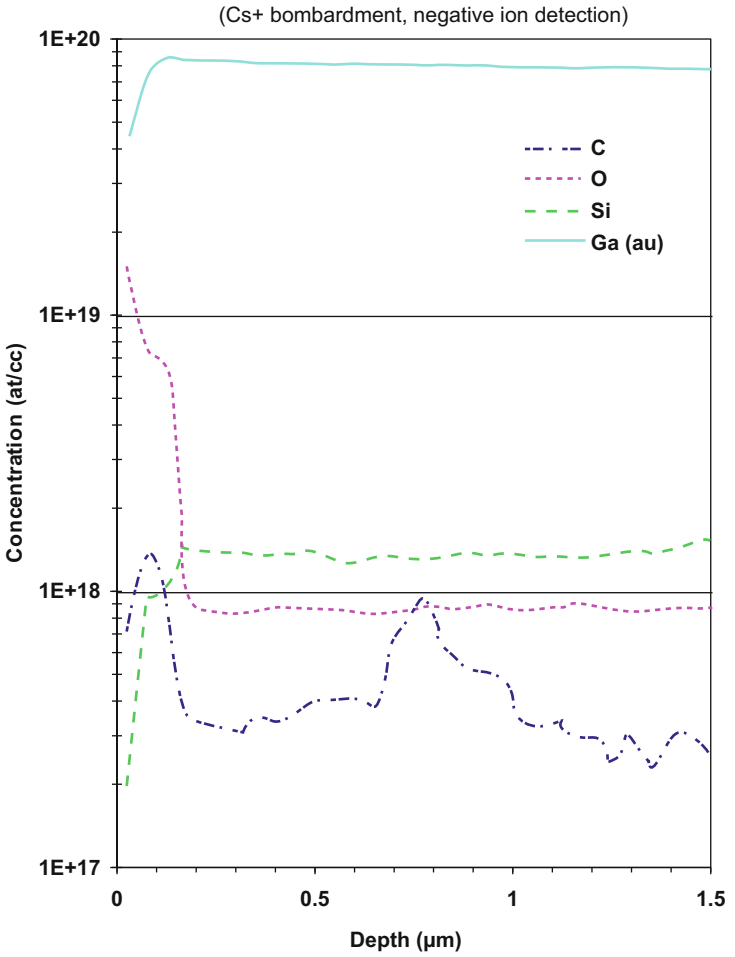


Fig. 18.11 A SIMS depth profile showing the concentration of impurities in a GaN sample. The impact energy was 15.5 keV at oblique incidence and the detected area was 33 μm in diameter

$$E_1 = \left\{ \frac{\left(M^2 - M_0^2 \sin^2\theta\right)^{1/2} + M_0 \cos\theta}{M_0 + M} \right\}^2 E_0 \tag{18.8}$$

where θ is the scattering angle. At a fixed value of M_0 and θ, E_1 depends only on the atomic weight of the target atom. Therefore, E_1 will be different for different targets, and by detecting this energy, one can distinguish between different atoms. This technique can be applied to multilayer samples as well. In this case not only the energy of the scattered beam but its intensity will also be affected by numerous scatterings inside the sample. In this case, top layers will have higher intensity scattered beams than the underlying layers.

18.2.8 Scanning Probe Microscopy (SPM)

Scanning probe microscopy (SPM) is a useful method for the study of the surface morphology. This method employs the concept of scanning an extremely sharp tip (3~50 nm radius of curvature) across the object surface. The tip is mounted on a flexible cantilever, allowing the tip to follow the surface profile (Fig. 18.12). When the tip moves in the proximity of the object under investigation, forces of interaction between the tip and the surface influence the movement of the cantilever. These movements are detected by selective sensors.

There are three major types of SPM:

- Atomic force microscopy (AFM) measures the interaction force between the tip and the surface. The tip may be dragged across the surface or may vibrate as it moves. The interaction force will depend on the nature of the sample, the probe tip, and the distance between them.
- Scanning tunneling microscopy (STM) measures a weak electrical current flowing between tip and sample as they are held a very short distance apart.
- Near-field scanning optical microscopy (NSOM) scans a very small light source very close to the sample. Detection of this light energy forms the image. NSOM can provide resolution below that of the conventional light microscope.

Essential to the system is a piezoelectric tube (Fig. 18.13). It consists of a piezo material inserted inside a hollow tube. Pairs of electrodes on the inner and outer walls are placed on either side of the tube. When suitable voltage differences are

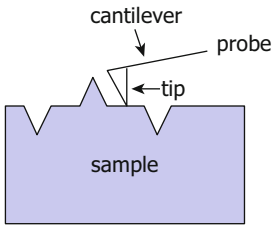


Fig. 18.12 Schematic of an AFM tip scanning over the surface of a sample

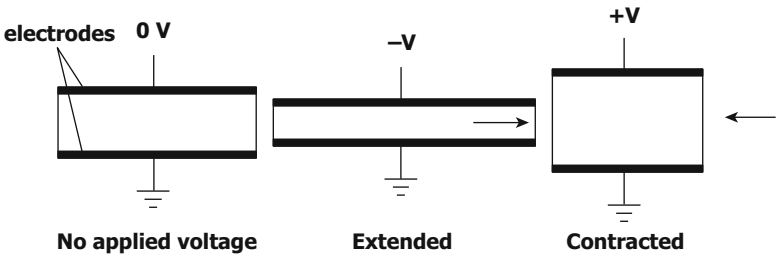


Fig. 18.13 Reaction of a piezo material to applied bias

applied to these electrodes, one side of the tube expands, and the other side contracts. This results in a bending of the tube; hence if one end is fixed, the other end moves, resulting in the scanning motion. Two sets of electrodes, 90 degrees apart, allow motion in the *x*-*y* plane. A further pair of electrodes extending around the entire circumference of the tube cause an entire section of the tube to expand or contract, resulting in the free end of the tube moving parallel to the tube axis (the *z*-axis). The combination of all three sets of electrodes allows movement of the free end of the tube to be controlled very precisely in all three axes. For surface mapping applications, the feedback provided by the probe and detector is used to keep the probe at a constant distance from the surface (*z*-direction), while it is free to move across the surface (*x*- and *y*-directions). This is accomplished by applying a voltage to the piezoelectric tube. This voltage is proportional to the probe's movement in *z*-direction which is then used to generate the surface topology.

The AFM is capable of reconstructing the surface morphology of the materials with atomic scale precision. An example of a three-dimensional image of the surface of InAs quantum dots grown on GaAs/InP is shown in Fig. 18.14.

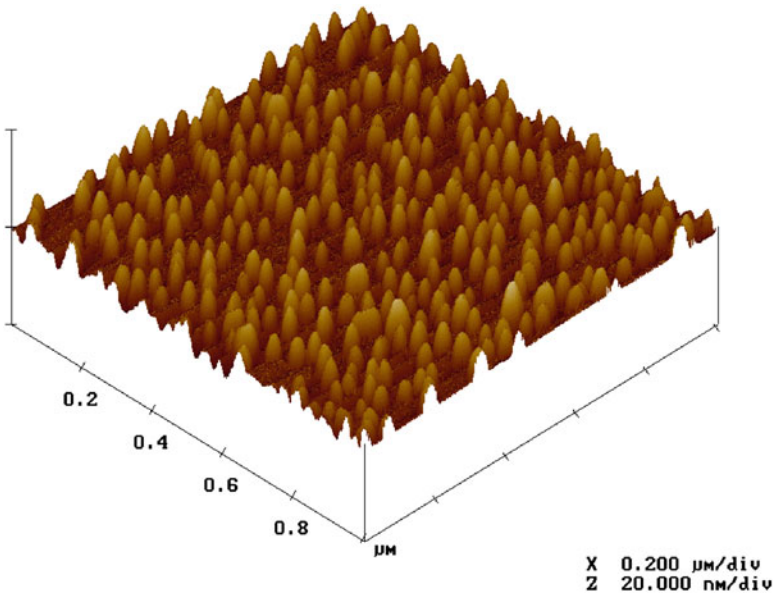


Fig. 18.14 A 3D AFM image of the surface of a sample consisting of InAs quantum dots grown on top of a GaAs/InP substrate

18.3 Optical Characterization Techniques

18.3.1 Photoluminescence Spectroscopy

Photoluminescence (PL) spectroscopy is a nondestructive method of probing the electrical properties of materials. Light is focused onto the sample where it is absorbed in a process called "photoexcitation." As a result of the excess energy caused by photoexcitation, electrons jump to permissible excited states. When these electrons move back to their equilibrium states, the excess energy is released through emission of light with energy equal to the energy difference between the equilibrium and excited states. This emitted light is then focused and collected by a photon detector through a spectrometer. A PL spectrum for an AlGaN sample is shown in Fig. 18.15. Many useful information can be extracted out of PL spectra:

- Bandgap determination: The most common radiative transition in semiconductors is between the states in the conduction and valence bands, which equals to the energy gap of the semiconductor.
- Impurity levels and defect detection: Radiative transitions in semiconductors involve localized defect levels. The photoluminescence energy associated with these levels can be used to identify specific defects.
- Recombination mechanisms: When the electrons return to their equilibrium states, also known as "recombination," both radiative and nonradiative processes can occur. The intensity of the PL peak and its dependence on the level of photoexcitation and temperature is directly related to the dominant recombination process.

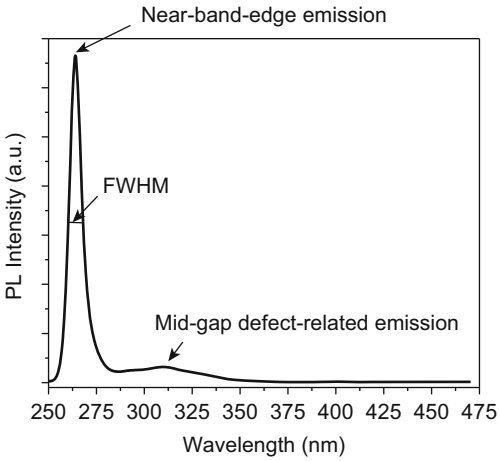


Fig. 18.15 Photoluminescence spectrum of an AlGaN sample. Shown on the graph are the near-band-edge emission peak and a defect-related emission peak

- Material quality: The intensity and the line width (FWHM) of a PL spectrum are representative of the quality of the material. Additionally, the presence of defect-related peaks is indicative of imperfections in the epitaxial layer.

18.3.2 Cathodoluminescence Spectroscopy

Cathodoluminescence (CL) spectroscopy is similar to PL in almost every aspect, except for the radiation source. In CL, electrons are used to excite the sample instead of photons in the PL case. The electron source can be the focused beam used in SEMs. Similar to PL spectra, CL spectra contain many useful information such as the ones listed in the previous subsection.

18.3.3 Reflectance Measurement

Any light incident upon any medium undergoes partial transmission, absorption, and reflection. The reflected part of the light can be collected and measured against a reference sample, typically a near-ideal mirror, to obtain the reflectivity. Reflectance is defined as the ratio of the reflected to incident light, given by Fresnel equations Eq. (10.22) as:

$$R = \left|\frac{E_\mathrm{r}}{E_\mathrm{i}}\right|^2 = \left(\frac{\bar{n}-1}{\bar{n}+1}\right)^2 \tag{18.9}$$

where E_r and E_i are the energy of the reflected and incident light, respectively, and $\bar{n}$ is the refractive index of the medium.

18.3.4 Absorbance Measurement

A visible/UV light beam is incident upon the sample under study and a reference sample simultaneously. The transmitted light out of the other face of the sample is collected by a photodetector through a spectrometer, and its intensity relative to the reference sample is plotted as a function of wavelength. This way one can determine the transmittance or absorbance of the sample under study as a function of wavelength. This method is especially useful for obtaining the absorption edge (cutoff wavelength) associated with the material. The band-to-band absorption in a semiconductor (see Chap. 10) gives the following relationship between the absorption coefficient α (see Eq. 10.81), the light energy E, and the bandgap energy E_g:

$$\alpha \propto \sqrt{E - E_\mathrm{g}} \tag{18.10}$$

18.3.5 Ellipsometry

Ellipsometry measures the change in the polarization state of light reflected from the surface of a sample. The measured parameters are the amplitude ratio (tan Ψ) and the phase difference (Δ) of the two components of reflected light. These values are related to the ratio of Fresnel reflection coefficients, R_p and R_s, for p- and s-polarized light, respectively:

$$\tan(\Psi)e^{i\Delta} = \frac{R_p}{R_s} \tag{18.11}$$

This simple fundamental equation of ellipsometry relates refractive indices of the film, and the substrate, film thickness, and phase changes during reflection at the film interfaces.

In Fig. 18.16, a linearly polarized input beam is converted to an elliptically polarized reflected beam. For any angle of incidence greater than 0° and less than 90°, *p*-polarized and *s*-polarized lights will be reflected differently.

The ellipsometry apparatus can also be used to measure transmission and reflection of samples. In this mode, the transmission (T) and reflection (R) values are determined via:

$$T = \frac{I_t}{I_i} \text{ and } R = \frac{I_r}{I_i} \tag{18.12}$$

where I_i, I_t, and I_r are the intensities of the incident, transmitted, and reflected lights, respectively.

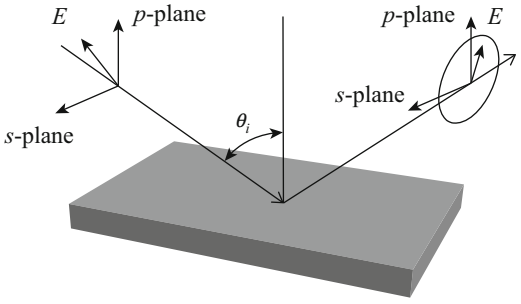


Fig. 18.16 Schematic of the geometry of an ellipsometry measurement. The coordinate system used to describe the ellipse of polarization is the p-s coordinate system. The s-direction is taken to be perpendicular to the direction of propagation and parallel to the sample surface. The p-direction is taken to be perpendicular to the direction of propagation and contained in the plane of incidence

18.3.6 Raman Spectroscopy

When photons are incident upon a medium, they get scattered either elastically (Rayleigh scattering) or inelastically (Raman scattering). In Rayleigh scattering, the energy of the emitted photon is the same as the incident photon. On the other hand, in Raman scattering, the energies of the scattered and incident photons are different. The energy change is depicted in Fig. 18.17, where an incoming photon either creates a phonon and is remitted at a lower energy (anti-Stokes scattering) or annihilates a phonon and is remitted at a higher energy (Stokes scattering). The inelastically scattered light can be collected, and information about the energy levels within the medium can be deduced from the energy change in the light.

A monochromatic light source, usually an argon ion laser, is used to excite the sample, and a spectrometer/PMT set is used to detect the scattered light. An example of a Raman spectrum is schematically shown in Fig. 18.18.

18.3.7 Fourier Transform Spectroscopy

A Fourier transform spectrometer is a Michelson interferometer with a movable mirror. By scanning the movable mirror over some distance, an interference pattern is produced that encodes the spectrum of the source (in fact, it turns out to be its

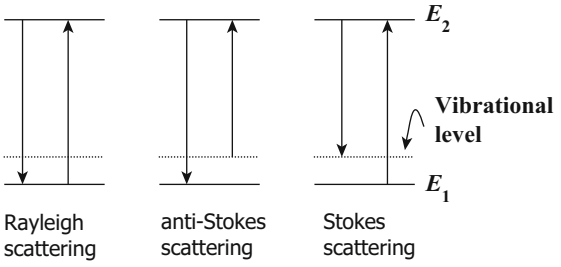


Fig. 18.17 Schematic depiction of various scattering processes within a medium. The incident photon energies are marked by the right-hand-side arrows

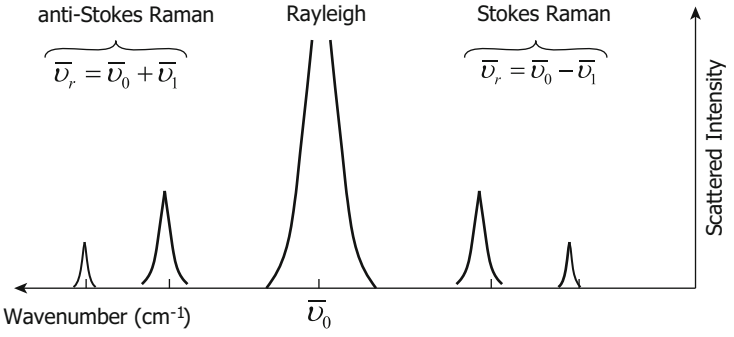


Fig. 18.18 An example of a Raman spectrum representing Rayleigh, Stokes, and anti-Stokes Raman peaks

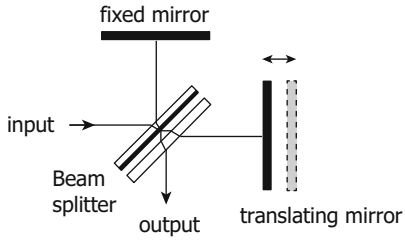


Fig. 18.19 Schematic cross section of a Michelson interferometer

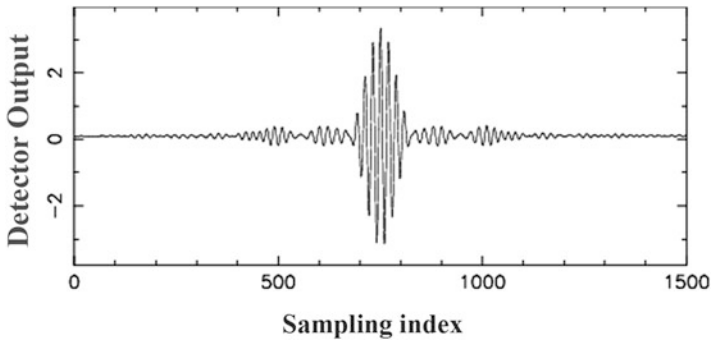


Fig. 18.20 A typical interferogram

Fourier transform). The Michelson interferometer consists of a beam splitter, a fixed mirror, and a mirror that moves back and forth as shown in Fig. 18.19. The input signal is split into two different optical paths, after which they add into the output signal. When the two mirrors are equidistant from the beam splitter, there is constructive interference for a given wavelength, and the output signal is very high. However, when the translating mirror is moving, its separation from the beam splitter varies, and the difference in distance that the two split beams of light have to flow through is called the optical path difference (OPD).

For incident light with a single wavelength, λ, on the input to the beam splitter, the output will have sinusoidal behavior with minima occurring when the OPD is an odd multiple of $\lambda/2$ (destructive interference). For a broadband incident light source, such as the luminescence from a semiconductor, the output intensity is more complicated as shown in Fig. 18.20. When the OPD is equal to zero, all spectral components interfere constructively; therefore, the absolute maximum of the interferogram, also called the center burst, is generated at that position. As the OPD increases, two different wavelengths will not reach a maximum output at the same time, giving us a complex looking oscillatory signal with decreasing amplitude, called the interferogram. It should be noted that when the wavelength of incident light is in the infrared region, this technique is called Fourier transform infrared (FTIR) spectroscopy.

The analog signal of the detector is digitized during the scan using A/D conversion running typically at frequencies up to 120 KHz with a numerical depth of 16 bits. In order to enhance the signal-to-noise ratio, some hundred scans are added

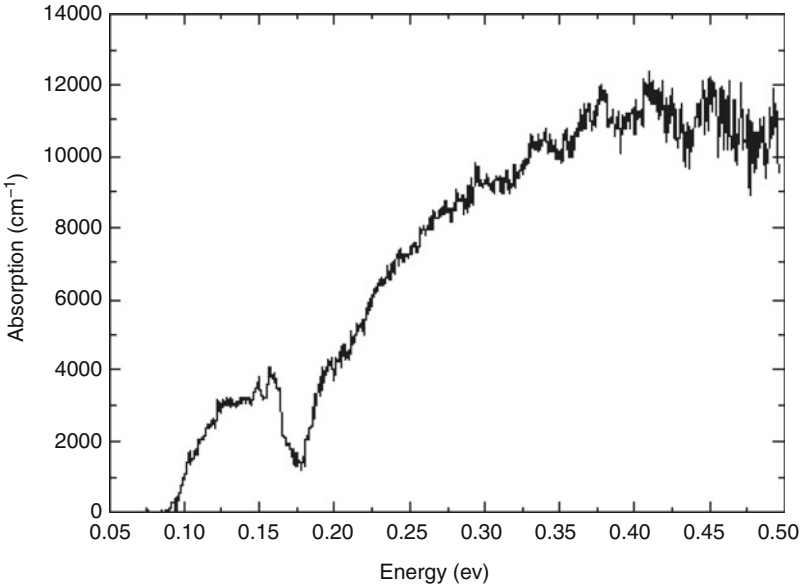


Fig. 18.21 Absorption spectrum for a semiconductor photodetector structure taken by a Fourier transform infrared (FTIR) system

coherently to build up the final interferogram. Once an interferogram is collected, it needs to be translated into an emission spectrum. The process of conversion is through the fast Fourier transform algorithm, which converts the time domain back into the frequency (or wavelength) domain. A typical example of an FTIR spectrum is shown in Fig. 18.21 illustrating the absorption of a semiconductor photodetector structure as a function of energy.

Normally, interferometric spectra are in units of wavenumber. The relationship between wavenumber and wavelength is:

$$\upsilon\left(\mathrm{cm}^{-1}\right) = \frac{10000}{\lambda(\mu m)} \tag{18.13}$$

Therefore, it would be easy to convert wavenumber to other useful units such as wavelength or energy, as is the case in Fig. 18.21.

18.4 Electrical Characterization Techniques

18.4.1 Resistivity

Using sheet resistivity measurement techniques (i.e., the four-point probe technique or the van der Pauw method), one can determine the sheet resistivity, ρ_s (and if the layer thickness is known, the resistivity, ρ), of a semiconductor layer. The

concentration of dopants can also be obtained from sheet resistivity measurements if the value of mobility is known (Eq. 8.8). Usually the carrier mobilities of some of the more established semiconductors, such as silicon, are known, and one can use those values to determine the carrier concentration from resistivity values. However, the type of doping (*n*-type or *p*-type) cannot be deduced from resistivity measurements. This technique is also useful when the carrier concentration varies as a function of depth. In this case, the resistivity will be:

$$\rho(z) = [N(z)e\mu(N)]^{-1} \tag{18.14}$$

where $N(z)$ is the carrier concentration as a function of depth and $\mu(N)$ is the carrier mobility as a function of carrier concentration. The measured sheet resistivity will be the weighted average given by:

$$\rho_s = \left[\int_0^t N(z)e\mu(N)dz\right]^{-1} \tag{18.15}$$

where t is the thickness of the layer.

18.4.2 Hall Effect

With Hall effect measurements, one can determine the concentration as well as the type of the dopants. In addition, the Hall mobility can be deduced from these measurements. Generally Hall effect measurement systems are capable of measuring low carrier concentrations, as low as 10^{14} cm^{-3}. The problems with Hall effect measurements are the rather difficult sample preparation (including contact preparation) and the errors that occur when the substrate is conductive. The reader is referred to Chap. 8 for a complete discussion on the Hall effect.

18.4.3 Capacitance Techniques

In capacitance techniques the charge storage capacity, or capacitance, is measured across a rectifying junction.

Capacitance-voltage (C-V) measurements use a time-varying voltage of variable frequency to determine the majority carrier concentration in the bulk of the device and/or energy levels of interface states that often exist between the surfaces of dissimilar materials. In order to determine the carrier concentration, usually a Schottky diode is built. The diode is then reverse biased and the value of capacitance is measured at each bias point. The carrier concentration can then be calculated as (refer to Chap. 9 for more discussion on junction capacitance):

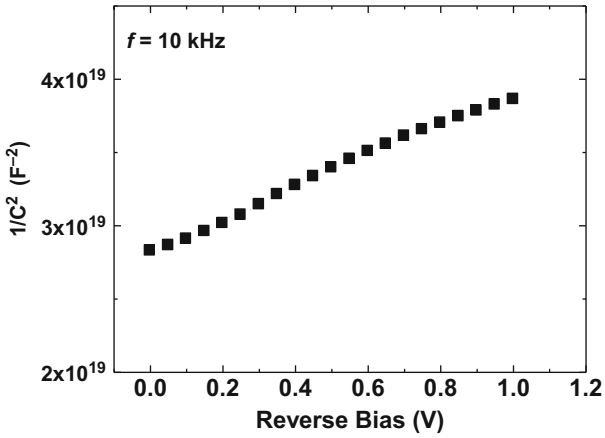


Fig. 18.22 Plot of C^{-2} vs. reverse bias for a p-type GaN sample. The measurements were taken at a frequency of 10 kHz

$$N = \frac{2}{\varepsilon\varepsilon_0 A^2}\left(\frac{1}{d\left(C^{-2}\right)/dV_r}\right) \tag{18.16}$$

where N is the carrier concentration (N_A for p-type, N_D for n-type), ε is the dielectric constant, A is the area of the diode, C is the capacitance, and V_r is the reverse bias. Figure 18.22 shows the plot of $\frac{1}{C^2}$ as a function of reverse bias for a p-type GaN sample. From the slope of the curve and the values of the dielectric constant and the diode area, the majority carrier concentration can be calculated.

Deep-level transient spectroscopy (DLTS) is another capacitance technique that examines the time-dependent flow of charge into and out of localized energy states associated with defects in the semiconductor. DLTS can thus determine many important defect-related properties, such as the nature of defects and their activation energies.

18.4.4 Electrochemical Capacitance-Voltage Profiling

Electrochemical capacitance-voltage (ECV) profiling is a measurement technique that allows one to determine doping level at various depths within a semiconductor structure.

Originally this technique was simply an extension of the CV measurement technique that calculates the average carrier concentration by measuring the capacitance across a Schottky barrier depletion region. In the modified approach, the sample is located inside an electrolyte that produces a well-defined electrochemical dissolution with the semiconductor material. This approach has led to the development of automated ECV profiling systems with nanometer etch depth resolution.

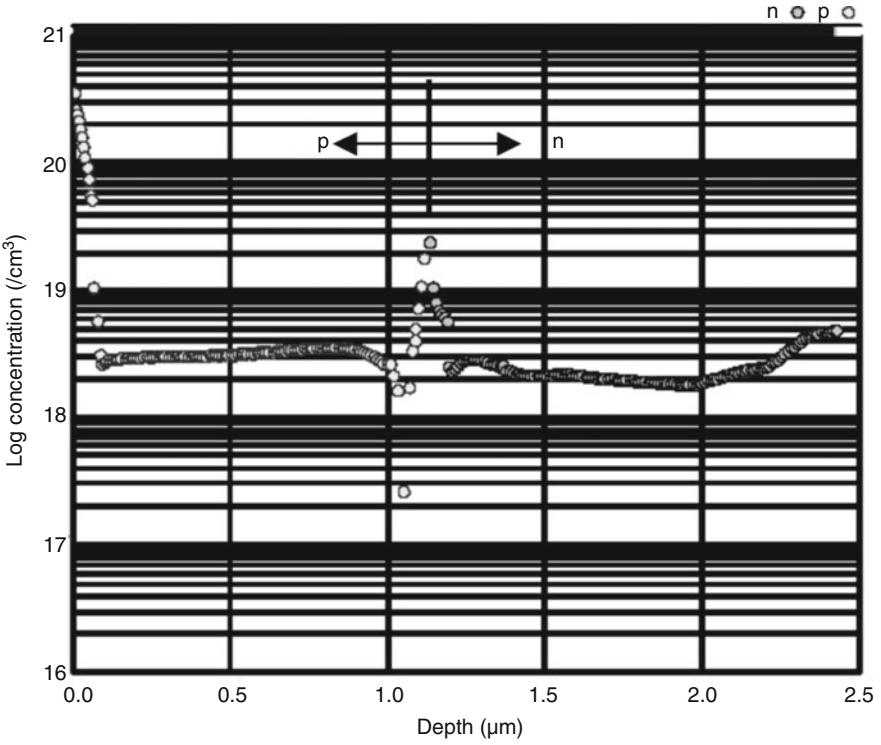


Fig. 18.23 A representative ECV profile showing the concentration of n-type and p-type dopants as a function of depth for a 980 nm laser diode structure

With the ECV profiling, it is not only possible to determine the type of doping (*n*-type, *p*-type) but also the concentration of the dopants in the range of 10^{13}–10^{21} cm^{-3}. An example of an ECV profile is shown in Fig. 18.23

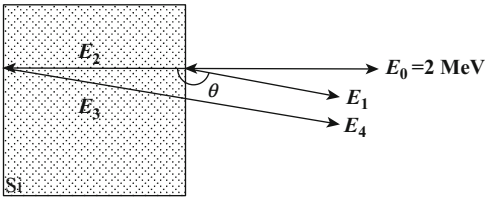
18.5 Summary

In this chapter we discussed several important semiconductor characterization techniques, covering structural, optical, and electrical properties of semiconductors. X-ray diffraction, electron microscopy (SEM and TEM), energy dispersive analysis using X-rays (EDX), Auger electron spectroscopy (AES), secondary-ion mass spectroscopy (SIMS), Rutherford backscattering (RBS), and scanning probe microscopy (SPM) were covered under structural characterization techniques. Optical characterization techniques included photoluminescence spectroscopy (PL), cathodoluminescence spectroscopy (CL), reflectance and absorbance measurements, ellipsometry, Raman spectroscopy, and Fourier transform spectroscopy. Finally, we briefly discussed some of the electrical characterization techniques such as resistivity

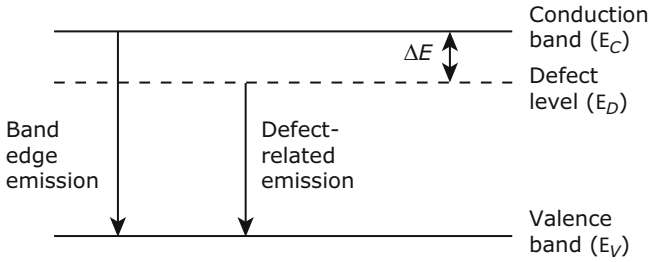
measurement, Hall effect measurement, capacitance techniques, and electrochemical capacitance-voltage (ECV) profiling. These characterization techniques are instrumental in understanding the most important properties of various semiconductors as building blocks of many useful electronic and optoelectronic devices.

Problems

1. The incident ion in an RBS measurement setup is $^4He^+$ at $E_0 = 3$ MeV. The angular position of the ion detector, θ, is chosen to be 170°. The backscattered beam from the surface of the sample under test has an energy of 2.5886 MeV. Determine which element of the periodic table the sample under test is made of.
2. In an RBS measurement setup, $^4He^+$ at $E_0 = 2$ MeV is used as incident ions. The scattering angle, θ, is 170 °. The incident ions impinge on a 100 nm thick silicon sample (atomic mass of Si equals 28.08). The majority of He ions penetrate below the surface where they lose their energy at a linear rate of 2 keV/nm. Determine the range of the backscattered energies from the sample ($\Delta E = E_4 - E_1$).



3. Estimate the acceptor concentration of the *p*-type GaN of Fig. 18.22. assuming a diode area of 400 μm × 150 μm and a dielectric constant of $\varepsilon = 10\varepsilon_0$.
4. Based on the SIMS spectrum of Fig. 18.11:
   (a) Estimate the thickness of the oxide layer that has formed on the surface.
   (b) Si is an *n*-type dopant in the GaN material system. What is the doping concentration away from the surface?
5. Based on the photoluminescence spectrum of Fig. 18.15:
   (a) Estimate the Al mole fraction (x) in the $Al_xGa_{1-x}N$ layer. Assume that Vegard's law holds for the calculation of the bandgap energy of the ternary $Al_xGa_{1-x}N$ from the binary compounds GaN ($E_g = 3.4$ eV) and AlN ($E_g = 6$ eV).
   (b) Assuming that the defect-related emission peak arises from the transitions from the valence band to a deep level, estimate how deep into the bandgap this deep level rests with respect to the conduction band edge ($\Delta E = E_C - E_D$).



6. In this chapter we introduced four measurement techniques that yield the impurity concentration in semiconductor layers, namely, SIMS, sheet resistivity (SR) measurements, Hall effect measurements, and ECV profiling. Complete the following table to compare these four techniques with respect to the stated application requirements.

Application requirement	SIMS	SR	Hall	ECV
Determination of doping concentration	✓	✓	✓	✓
Determination of doping type (n-type or p-type)				
Determination of the concentration of electrically activated dopants				
Easy sample preparation				
Determination of dopant concentration as a function of depth				
Non-destructive measurement				
Thickness of the layer may be unknown				

7. From the discussion of Rayleigh scattering, we recall that Rayleigh scattering is the elastic scattering of light off molecules that are smaller than the wavelength of that light. The intensity of the scattered light as a function of wavelength is given by:

$$I = I_o\left[\frac{8\pi^4 N\alpha^2}{\lambda^4 R^2}\right]\left(1 + \cos^2\theta\right)$$

Based on this formula, justify why the sky appears blue.

8. Do you think SEM and AFM are competing techniques or complementary techniques? Explain why.
9. Based on the TEM image provided in Fig. 18.7, estimate the lattice mismatch between AlN and sapphire.
10. When an X-ray beam impinges upon a sample, it gets partially transmitted, partially absorbed, and partially scattered (diffracted). The ratio of the intensity of the transmitted beam to that of the incident beam can be expressed as: $\frac{I_T}{I_0} = e^{-\alpha x}$ where α is a constant and x is the thickness of the sample. We know that if the thickness of a sample is doubled, it means that the number of crystallographic planes that cause diffraction from a transmitted beam has been doubled. Based on this, propose a formula that describes the intensity of the diffracted beam versus the incident beam. At what thickness is this intensity maximum? What percentage of light will be transmitted at this optimum thickness?

References

Ohring M (1992) The materials science of thin films. Academic Press, San Diego
Williams DB, Carter CB (1996) Transmission electron microscopy. Plenum Press, New York

Further Reading

Long DA (1977) Raman spectroscopy. McGraw-Hill, New York
Perkowitz S (1993) Optical characterization of semiconductors: infrared, Raman, and photoluminescence spectroscopy. Academic Press, London
Razeghi M (1989) The MOCVD challenge volume 1: a survey of GaInAsP-InP for photonic and electronic applications. Adam Hilger, Bristol
Razeghi M (1995) The MOCVD challenge volume 2: a survey of GaInAsP-GaAs for photonic and electronic device applications. Institute of Physics, Bristol, pp 21–29
Stradling RA, Klipstein PC (1990) Growth and characterization of semiconductors. Adam Hilger, New York
Warren BE (1990) X-ray diffraction. Dover Publications, New York

Defects

19

19.1 Introduction

An ideal crystalline solid has a periodic structure that is based on the chemical properties of its constituent atoms (see Chap. 3). However, real crystals are not perfect. They always have imperfections such as extra/missing atoms or impurities, which are called defects.

The periodicity characterizes the crystals as we learned in previous chapters. For example, the periodic potential of the lattice modulates the wavefunction, and we can establish relationships between the energy and wavevector using the Bloch theorem as shown by the Kronig-Penney model (Chap. 5). The existence of defects perturbs the potential of the lattice, and this modifies the band diagram in the crystals.

While many properties of crystalline systems depend upon the periodic lattice arrangement, many additional properties can be manipulated by adding defects or dopants to the crystal. These properties enable us to fabricate various devices in the modern world of semiconductor technology. On the other hand, unintentionally introduced defects can also have a profound impact on the properties of materials or on the performance of these devices. Therefore, it is a challenging goal to have precise control of defects in crystals.

The defects can determine the color of the crystal and its electric conductivity, and they can also introduce modifications in the lattice vibrations. For example, silicon becomes *p*-type with boron doping. Al_2O_3 has red color as a ruby when a small amount of Cr^{3+} substitutes Al^{3+}, but Al_2O_3 has blue color as a sapphire when a small amount of Ti^{3+} is substituted for Al^{3+}.

In this chapter, we will discuss how defects are introduced in crystals and the possible reasons or sources of such imperfections, which may be roughly summarized as follows:

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7_19

(i) *Defects from fundamental physical laws*
There are defects that must exist due to fundamental physical laws. One example is a vacancy. At any finite temperature, the atoms undergo a degree of vibrational displacements. As the temperature is raised, the displacements may become so large that atoms are permanently moved from their normal sites. These atoms leave their sites and vacancies are formed.

(ii) *Defects from natural minerals*
Materials are never 100% pure. Therefore, all crystals have certain foreign atoms; impurities as defects. Silicon wafers used in modern semiconductor technology are purified to a very high degree (better than 99.999999%).

(iii) *Defects from crystal growth* (see Chap. 17 for details)
Intrinsic defects can be introduced during crystal growth. For example, typical concentrations of intrinsic defects in Si are on the order of 10^{13}–10^{14} cm^{-3}. Extrinsic defects (impurities) can also be introduced in the crystallization process. The species of the impurities depends on the growth method and on the constituent materials of the growth system.

(iv) *Defects from strain*
Deformation of metals or any strain added to crystals generates defects (mainly dislocations). Especially in semiconductor technology, the defects caused by strain are of great interest for heteroepitaxial thin film growth. For example, semiconductor lasers and integrated-optics devices are usually designed from multilayer structures which have similar lattice constant because the mismatch of lattice parameters accumulates strain and results in the creation of undesirable defects. The defects caused by lattice mismatch are efficient non-radiative recombination channels and therefore should be avoided since they degrade the performance of optical devices. However, the recent increasing demand for wide bandgap materials such as GaN has confronted the growers with exactly this difficulty. Since GaN has no readily available native lattice matched substrate, and the lattice mismatch depends on the substrate, these materials cannot be obtained without lattice mismatch. In addition, there also exist devices which positively make use of the effect of strain, such as high-electron-mobility transistors (HEMT) and self-organized strain-relaxed islands (quantum dots) made in the Stranski-Krastanov growth mode (Chap. 17). For these applications, the defects caused by strain constitute the active layer.

There are several categorizations of defects. One of the common classifications is based on the dimension of the defect structure. Defects may be classified into four groups: point defects (0D), line defects (1D), planar defects (2D), and volume defects (3D). Table 19.1 displays examples of these four types of defects.

Table 19.1 Table of dislocation dimension classifications

Dimension	Examples
0D: Point defects	Vacancies, self-interstitials, impurities
1D: Line defects	Edge dislocations, screw dislocations, mixed dislocations
2D: Planar defects	Stacking faults, grain boundaries, twin boundaries, interphase boundaries, external surfaces
3D: Volume defects	Precipitates, voids

19.2 Point Defects

Point defects, or 0-dimensional defects, refer to missing, additional, or misplaced atoms within the crystalline lattice. Figure 19.1 shows examples of substitutional, interstitial, and vacancy point defects, each of which will be discussed in more detail in the following sections.

19.2.1 Intrinsic Point Defects

The presence of intrinsic point defects is related to the nature of the atom. Atoms in a solid are subject to thermal vibrations at any temperature. The average amplitude of the atomic displacements increases with increasing temperature. Therefore, it is easy to imagine a localized area within the crystal where the vibrations are intense enough to cause a single atom to jump to a different location, either to the surface of the crystal or to an intermediate or interstitial position within the crystal. If the atom moves to the surface of the crystal, a Schottky defect is said to have formed, leaving a vacancy as the defect. However, if the atom jumps to an interstitial position within the crystal lattice, it is said to have formed a Frenkel defect, creating both a vacancy and a self-interstitial. A vacancy is a missing atom within the crystal lattice. A self-interstitial is an atom of the same type as the bulk material that is located at a non-lattice site. A Schottky defect is shown schematically in Fig. 19.2a, while a Frenkel defect is shown schematically in Fig. 19.2b.

It has been shown experimentally that at thermal equilibrium, all crystals contain intrinsic point defects. This leads to the conclusion that the imperfect crystal has a lower free energy than a perfect crystal. From thermodynamics, we know that the change in the free energy of a system, ΔG, is related to the changes in enthalpy, ΔH, and entropy, ΔS, as shown in Eq. (19.1), where T is absolute temperature:

$$\Delta G = \Delta H - T\Delta S \tag{19.1}$$

The energy to form a defect, E_D, is a positive contribution to the enthalpy term, thus *increasing* the free energy of the system. However, the creation of the defect increases

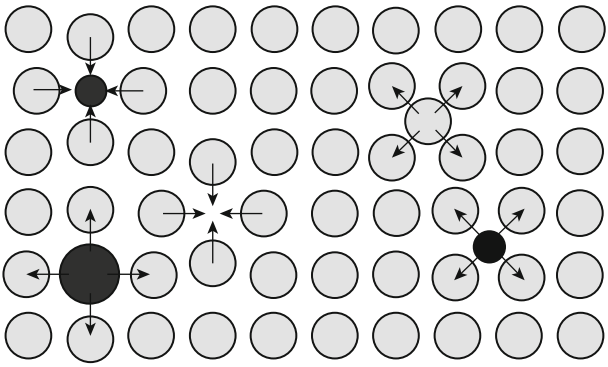


Fig. 19.1 Examples of point defects

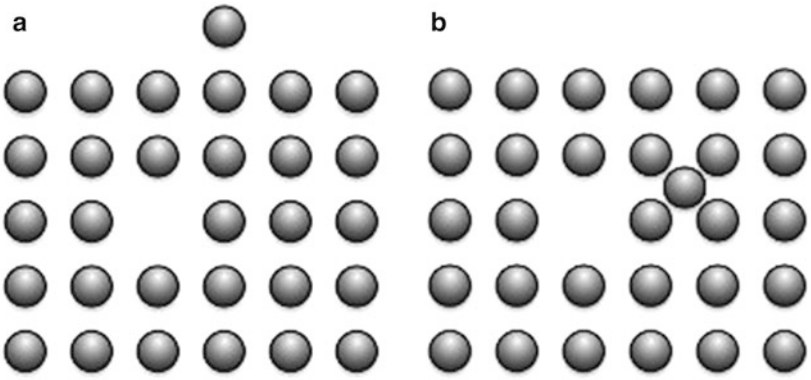


Fig. 19.2 Schematic diagrams of a: (**a**) Schottky defect and (**b**) Frenkel defect

the disorder of the crystal, thus increasing the entropy of the system and causing a *decrease* in the free energy of the system. The balance of these two factors leads to an equilibrium number of defects naturally occurring within the crystalline lattice. Through calculating the minimum free energy condition as a function of temperature, Boltzmann determined that the equilibrium number of defects, n_e, can be written according to Eq. (19.2), where N is the number of atoms in the crystal, A is a constant often taken as unity, T is the absolute temperature, and k_b is the Boltzmann constant. By dividing n_e by N, the equilibrium concentration of defects, n_e, may be found.

$$n_e = NA\exp\left(\frac{-E_D}{k_b T}\right) \tag{19.2}$$

One key process that affects both semiconductor device performance and some fabrication techniques is chemical diffusion. Chemical diffusion occurs when atoms of the same type or a different type are able to move through the crystalline lattice over time. The presence of vacancies in a solid enhances the rate at which chemical

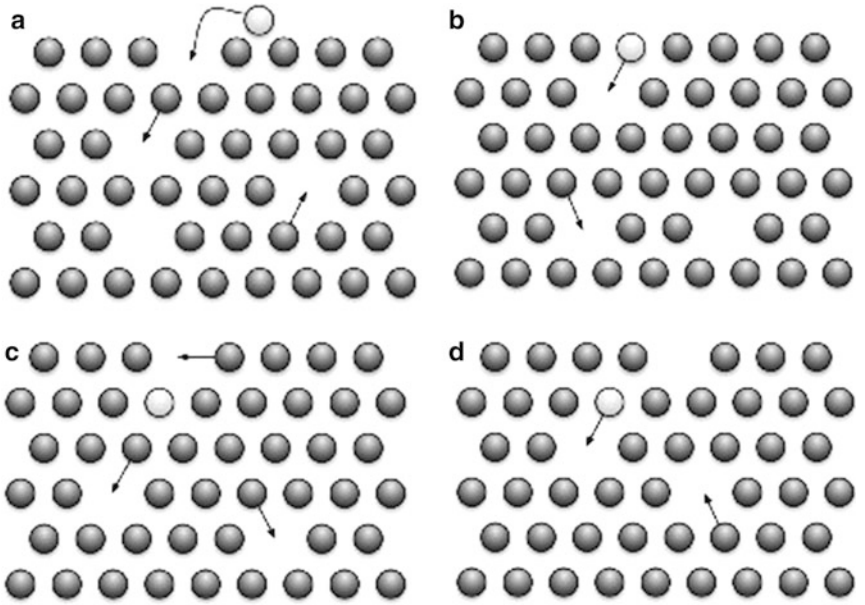


Fig. 19.3 Schematic of chemical diffusion showing how a foreign atom may diffuse into a crystal with time assisted by the presence of voids (increasing time from (**a**) to (**d**))

diffusion takes place. It is easy to imagine, for example, oxygen atoms diffusing from the surface of silicon into the silicon crystalline lattice through vacancies, as shown in Fig. 19.3.

Furthermore, it is also expected that at higher temperatures, when there are more vacancies in the network, the diffusion through the vacancy sites of the lattice takes place at a higher rate. The oxygen atom reaches a deeper site within the crystal more rapidly.

Another type of intrinsic point defect is an anti-site defect, shown in Fig. 19.3. An anti-site defect can occur when the crystalline lattice contains at least two kinds of atoms. Given enough energy, it is possible for two atoms to trade positions in the lattice. This is another diffusion mechanism, termed rotation about a midpoint.

19.2.2 Extrinsic Point Defects

Extrinsic point defects, shown schematically in Fig. 19.4, are caused by an outside source, such as growth conditions or processing factors. They are created when a foreign atom embeds itself within the crystal. If the atom is located on a lattice site, i.e., replacing the native atom, then it is called a substitutional impurity. The foreign atom may also be located at an interstitial site and is thus termed an interstitial impurity.

It is virtually impossible to control all environmental factors in order to have a 100% pure material, although for some applications this is highly desirable. The type

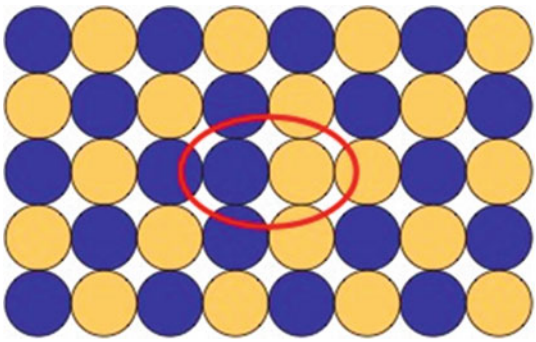


Fig. 19.4 Schematic diagram of an anti-site defect

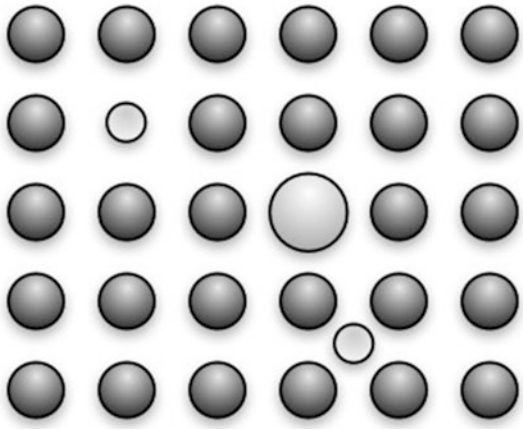


Fig. 19.5 Diagram of extrinsic point defects of substitutional impurities and an interstitial impurity

of the impurity depends on each growth method and the materials used in the system. For example, one of the major contaminations in MOCVD growth is carbon from group III sources. With respect to silicon technology, from the many possible impurities, it is the incorporation of metallic impurities that must be reduced to extremely low levels. This is because most metals have low solubility in silicon, and this results in metal silicides forming near the surface during device processing. Furthermore, many metals form deep traps in the energy bandgap of semiconductor materials, and this shortens the minority carrier lifetime considerably (Fig. 19.5).

There are also cases where impurities are desirable. In those cases, the challenge is the control of the type of impurity to be incorporated at well- defined lattice sites or specific regions within the crystal with precise concentration.

The most important application of extrinsic defects, especially with respect to semiconductors, is doping. While in many cases it is undesirable to have foreign atoms located within a crystal, doping purposely creates substitutional impurities in order to give the crystal certain properties. For example, GaN is doped with magnesium ions in order to create *p*-type GaN. Without achieving controlled doping, semiconductor devices would not exist. For more detailed information on doping, see Chap. 9.

Table 19.2 Impurity ionization energy (in meV) for several semiconductors (Wolfe et al. 1989)

Si		Ge		GaAs		GaP	
Li^+	32.81	Li^+	9.89	Si^+	5.854	Si^+	82.1
P^+	45.31	P^+	12.76	Ge^+	5.908	Ge^+	201.5
As^+	53.51	As^+	14.04	Sn^+	5.817	Sn^+	65.5
Sb^+	42.51	Sb^+	10.19	S^+	5.89	S^+	104.2
Bi^+	70.47	Bi^+	12.68	Se^+	5.808	Se^+	102.6
B^-	45	B^-	10.47	Te^+	5.892	Te^+	89.5
Al^-	57	Al^-	10.80	Be^-	30	Be^-	48.7
Ga^-	65	Ga^-	10.97	Mg^-	30	Mg^-	53.5
In^-	160	In^-	11.61	Zn^-	31.4	Zn^-	64
		Ti^-	13.10	Cd^-	35.4	Cd^-	96.5
				C^-	26.7	C^-	48
				Si^-	35.2	Si^-	203
				Ge^-	41.2	Ge^-	257
				Sn^-	171		

For doping to add carrier concentration or change the carrier type, impurities with shallow activation or ionization energies are used. For *p*-type silicon, boron is usually the preferred dopant, while phosphorus, arsenic, and antimony are used for *n*-type. Some of the activation energies are listed below in Table 19.2 (note: data about the most common dopants in Si, Ge, and GaAs was already listed in Table 19.1).

19.3 Line Defects

Line defects, or one-dimensional defects, refer exclusively to dislocations. Although there are two main types of dislocations, edge or screw, these two types typically combine to form several complicated mixed dislocations.

Edge dislocations may be described as an extra plane of atoms inserted into the crystalline lattice, causing a localized strain to be introduced into the lattice, as shown in Fig. 19.6.

Screw dislocations are formed when one side of the crystal undergoes a shear stress and is displaced at least one lattice plane, while the other side is held fixed. A schematic diagram of a screw dislocation is shown in Fig. 19.7.

Mixed dislocations are any combination of edge and screw dislocations and are the most typical ones that one finds in bulk crystals. An example of a simple mixed dislocation is shown in Fig. 19.8.

Burgers' vectors are used to classify and describe dislocations. In order to construct a Burgers' vector, a closed loop should be drawn around the dislocation by traveling the same amount of lattice points in all directions. If the loop does not close, it is surrounding a dislocation, and the vector that would close the circuit is the Burgers' vector. The starting point, the circuit direction, and the size of the loop are arbitrary. Independent of these factors, the Burgers' vector will always be perpendicular to the

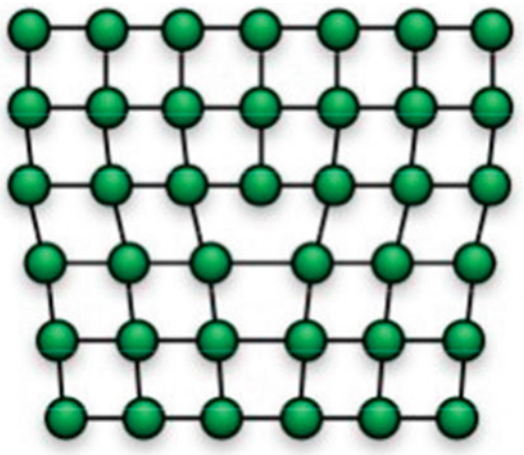


Fig. 19.6 Illustration of an edge dislocation

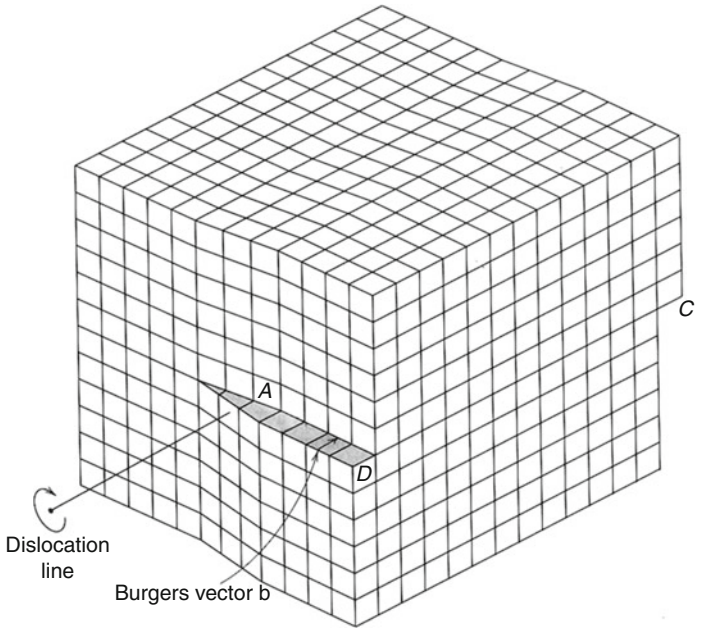


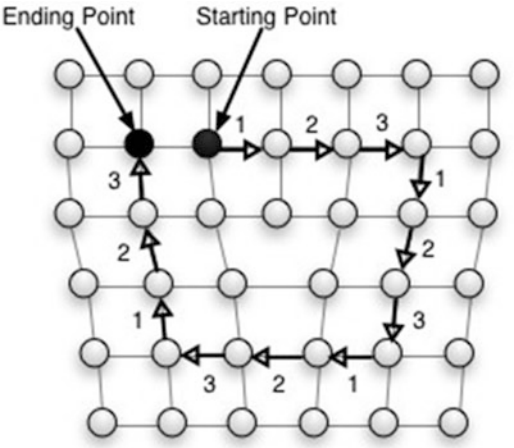
Fig. 19.7 Illustration of a screw dislocation. (Materials science and engineering: an introduction, Callister WD, p. 76, Fig. 4.4(a) Copyright © 2000 by John Wiley & Sons, Inc. Reprinted with permission of Wiley-Liss Inc., a subsidiary of John Wiley & Sons, Inc.)

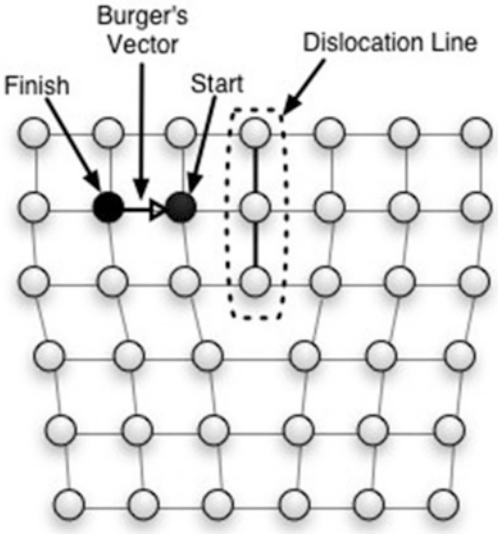
line of an edge dislocation and parallel to the line of a screw dislocation. It is often very complicated to find the Burgers' vector for a mixed dislocation.

Example

Q: Draw the Burgers' circuit to show that the Burgers' vector for an edge dislocation is perpendicular to the line of the dislocation.

A: Choose a starting point, a direction, and a side length that will be sure to enclose the edge dislocation. In the figure below, a clockwise direction and a side length of three were chosen. Then draw a vector from the end point of your circuit to the starting point of your circuit. This is the Burgers' vector.





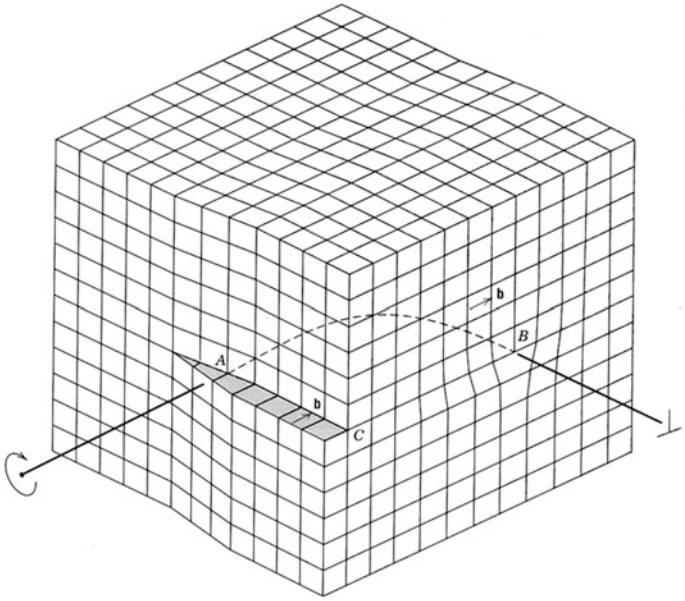


Fig. 19.8 Illustration of a mixed dislocation comprised of one edge dislocation and one screw dislocation. (Materials science and engineering: an introduction, Callister WD, p. 77, Fig. 4.5 (a) Copyright © 2000 by John Wiley & Sons, Inc. Reprinted with permission of Wiley-Liss Inc., a subsidiary of John Wiley & Sons, Inc.)

19.4 Planar Defects

Planar defects, or two-dimensional defects, refer to irregularities in the crystalline lattice that occur across a planar surface of the crystal. These may be due to an internal error in the crystal structure, or interfaces between two different materials, including interfaces with different phases of matter. Internal planar defects include stacking faults, twin boundaries, grain boundaries, and interphase boundaries, while external planar defects refer to surface defects caused by an interaction of the crystal with a gas or liquid environment.

Stacking faults occur when a single plane of atoms within the crystalline lattice is misoriented or out of order. For example, the cubic close-packed structure follows an ABCABC stacking order; however, an error in this order such as a stacking of ABCABABC produces a stacking fault. Figure 19.9 shows an example of a stacking fault.

Twin boundaries occur when a stacking fault reorients the rest of the crystal, forming a mirror plane within the crystal. For example, in the ABCABC stacking order of the cubic close-packed structure, a new stacking order of ABCABACBA would cause a twin boundary, where the center "B" plane would be a mirror plane. A schematic of a twin boundary is shown in Fig. 19.10 .

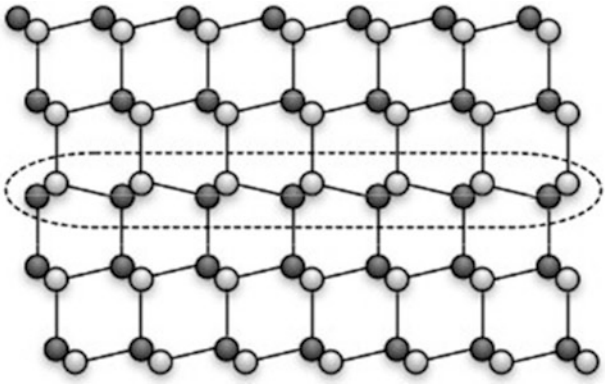


Fig. 19.9 Schematic diagram of a stacking fault

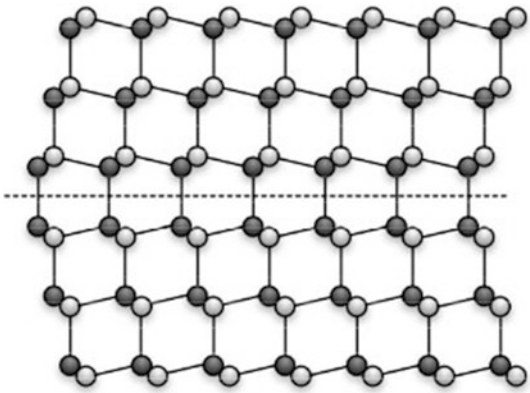


Fig. 19.10 Schematic diagram of a twin boundary

When two or more single crystals of different orientation meet, grain boundaries are formed. Two types of grain boundaries are pure tilt boundaries and pure twist boundaries. Pure tilt boundaries occur when the axis of rotation is parallel to the plane of the grain boundary, as shown in Fig. 19.10.

Pure twist boundaries, on the other hand, occur when the axis of rotation is perpendicular to the plane of the grain boundary, as shown in Figs. 19.11 and 19.12.

If the angle of rotation is small enough for these two cases, usually less than 10°–15°, the grain boundary is referred to as small angle. A small angle pure tilt boundary can be viewed as a series of parallel edge dislocations, while a small angle pure twist boundary may be viewed as an array of screw dislocations. The spacing between the dislocations, D, of low-angle grain boundaries is given in Eq. (19.3), where b is the magnitude of the Burgers' vector, which measures the degree of the misalignment introduced into the lattice due to one dislocation, and θ is the rotation angle.

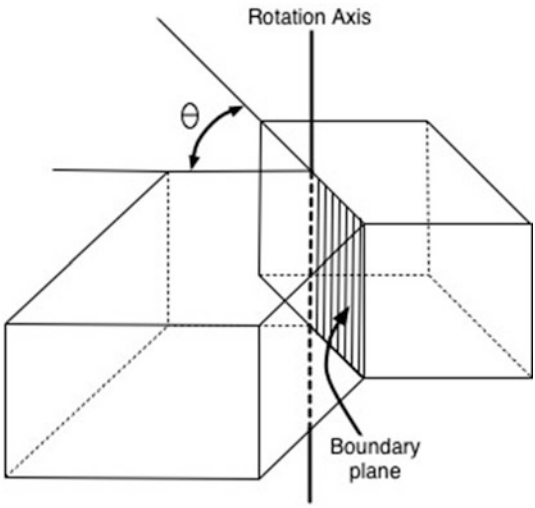


Fig. 19.11 Schematic diagram of a tilt boundary

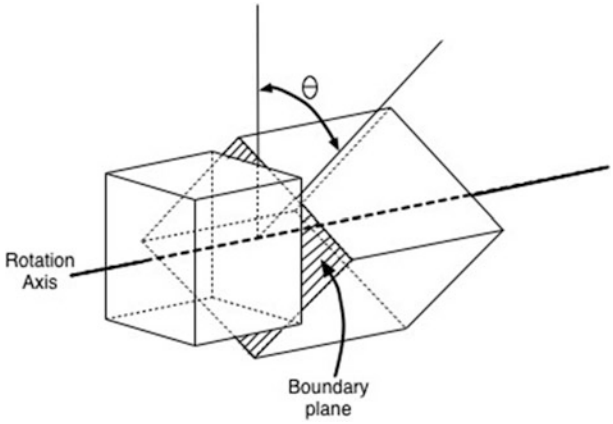


Fig. 19.12 Schematic diagram of a twist boundary

$$D = \frac{b}{\sin\theta} \approx \frac{b}{\theta} \tag{19.3}$$

Large-angle grain boundaries and combinations of twist and tilt boundaries lead to much more complicated structures for grain boundaries. Polycrystalline materials generally contain many grains of single crystalline material of random orientations with their neighbors. The size of the grains and the orientation between neighboring grains have an effect on properties of the polycrystalline material. For instance, a material with large grains and only a small misorientation between grains would have properties closer to a single crystalline material than a material with small, highly disordered grains.

Interphase boundaries occur when one crystalline material shares an interface with another crystalline material. Depending on the properties of each material, the interface will be either coherent, semi-coherent, or incoherent.

Coherent interphase boundaries will form when the two materials have similar geometries and a layer thickness less than the critical thickness for that material interface. The critical thickness, d_{crit}, is approximated by Eq. (19.4) where b is the magnitude of the Burgers' vector for a dislocation and f is the lattice mismatch between the two materials. Since the critical thickness is indirectly proportional to the lattice mismatch of the two materials, in order to have a coherent interface, it is necessary to have a small enough lattice mismatch in order to have a reasonable critical thickness (thicker than a few monolayers):

$$d_{crit} = \frac{b}{10 \cdot f} \tag{19.4}$$

While a small amount of strain may be introduced at a coherent boundary, no defects will be introduced due to the material change. A coherent boundary is shown in Fig. 19.13.

Semi-coherent interphase boundaries will form when the two materials have similar geometries but a larger lattice mismatch or the layer thickness exceeds the critical thickness. In this case, edge dislocations tend to form due to increased strain within the material. A semi-coherent boundary is shown in Fig. 19.14.

Incoherent interphase boundaries have a highly disordered structure that lack orientation relationships and have high energies. Little is known about the detailed structure of this type of interface.

External planar defects occur when the crystal periodicity is interrupted and bonds are broken, leading to dangling bonds. This occurs at the surface of the crystal and affects the outermost atomic layers or surface region. When this occurs, the atoms on the surface have a smaller coordination number, or number of nearest neighbors, than the atoms in the bulk crystal, and therefore have significantly different properties than the bulk crystal. The dangling bonds cause the surface to be more chemically and electrically active.

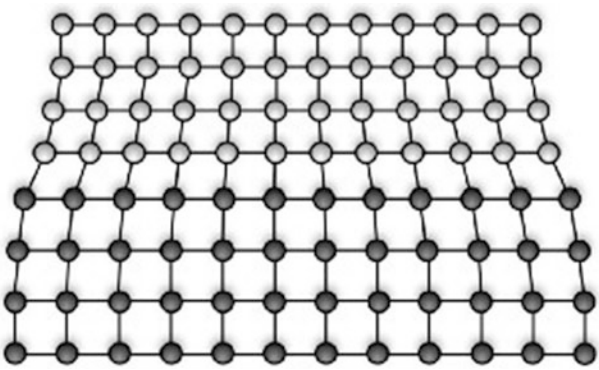


Fig. 19.13 Schematic of a coherent interphase boundary

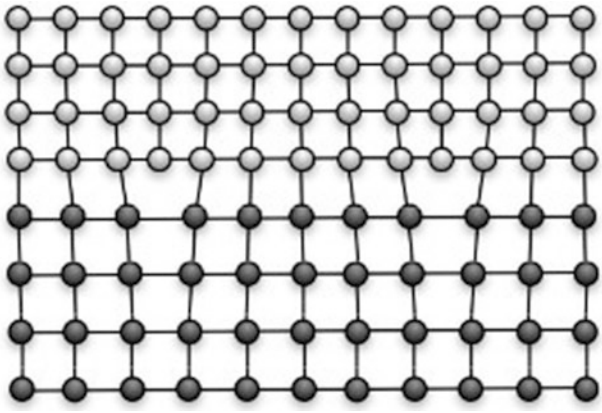


Fig. 19.14 Schematic of a semi-coherent interphase boundary

Since it takes energy to break the bonds, creating a surface takes energy, referred to as surface energy, which is always a positive amount. The surface wants to minimize its energy by reducing the number of dangling bonds, which it may do through surface relaxation or surface reconstruction. Surface relaxation is achieved by a change in the distance between the first and second layers of atoms at the surface. Typically, the distance is reduced, but there are a few cases where it is increased. Surface reconstruction occurs when the surface forms a different structure than the bulk structure. The silicon (001) surface relies on surface reconstruction in order to minimize its surface energy.

19.5 Volume Defects

Volume defects, also known as bulk defects, are clusters of point defects. Clusters of defects are produced when the crystal become supersaturated.

Each point defect introduced into a crystal has a certain level of solubility, which defines the maximum concentration of the impurity in the host crystal. In general, solubility is temperature dependent and decreases as the crystal is cooled down. When the concentrations of defects exceed their solubility limit or the crystal is cooled down after it gets saturated, it becomes supersaturated with that defect. The crystal under a supersaturated condition tries to achieve an equilibrium condition by condensing the excess defects into clusters with different phase regions.

Clusters of vacancies forming small regions where there are no atoms are called voids. High concentration of point defects in semiconductors results in formation of microvoids. The aggregation of vacancies is increasingly harmful to device performance as the size shrinking of devices continues in Si wafers Fig. 19.15 shows an SEM image of voids in AlGaN.

Clusters of foreign atoms forming small regions of different phase are often called precipitates. For example, Zn in InP at a doping level exceeding 1×10^{18} cm^{-3}

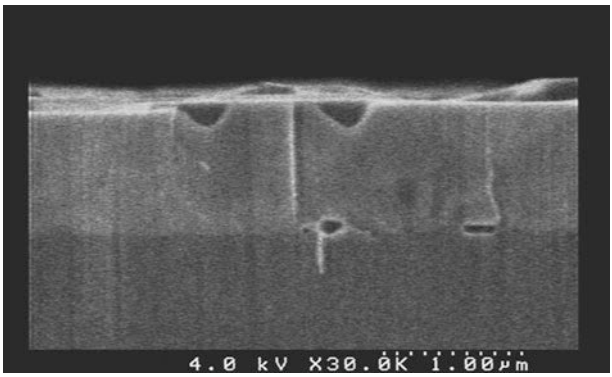


Fig. 19.15 SEM image of voids in AlGaN

forms precipitates. Another example is precipitates in silicon which occurs during the processing of wafers into integrated circuits. There are two foreign particle formation mechanisms: precipitates and inclusion incorporation. Precipitates are formed due to the retrograde solubility of native point defects. When the grown crystal is cooling down, the solidus line is crossed, and nucleation of the second phase takes place. In contrast to precipitates, inclusions are formed by capturing melt solution droplet from the diffusion boundary layer adjacent to the growing interface and enriched by the rejected excess component.

19.6 Defect Characterization

Characterization and analysis of defects is one of the biggest experimental challenges. There are conventional characterization methods to examine the overall quality or electrical features of the material such as Hall measurement and x-ray measurement (see Chap. 18). However, observing and identifying the type of each defect and the status in the material or devices is not easy because the defects are usually of atomic size unless they aggregate and form clusters.

When the defects are revealed by special etching techniques, they can be observed by optical microscopy. This method is called preferential etching. The basic idea of the method is to make defects visible in a microscope by marking the surface with small pits or grooves. This happens due to the differing physical and chemical properties near the defects. The surface is polished and etched with proper etching solutions that dissolve the material much more quickly around defects than in perfect regions.

Scanning electron microscope (SEM) has been used for observing large defects in devices in research and industry. For smaller features, transmission electron microscope (TEM) is now a better choice. Scanning probe microscope (SPM) and atomic force microscope (AFM) are capable of imaging single atoms. There are also several analytical methods for detecting impurities such as atomic absorption spectroscopy

(AAS), spark source mass spectrometry (SSMS), secondary ion mass spectrometry, and local mode infrared absorption.

19.7 Defects Generated During Semiconductor Crystal Growth

As previously mentioned, intrinsic defects will always exist at temperatures above the absolute zero. In reality, however, the actual defect concentrations in crystals are much higher than the equilibrium values at room temperature. This is because the finite defect diffusion rate leads to the freezing-in of a large fraction of the high-temperature defects produced as the crystal cools down. Therefore, pulling rate and cooling rate from the melting point are important parameters for crystal growth.

The development of crystal growth technology has been motivated by two major goals: achieve higher quality of bulk crystals and larger wafer diameters. Higher quality is necessary because as device sizes continue to shrink, the presence of defects in crystals becomes more significant. In particular, the aggregation of vacancies which results in the formation of microvoids is increasingly harmful to device performance. Large-diameter wafer development is driven by the demand of cost reduction in the device industry, since larger wafer diameter leads to higher throughput.

The growth of compound semiconductor single crystals is more complicated and less studied compared to Si, for instance. In III-V and II-VI semiconductors, the intrinsic point defect concentration is even greater than the intrinsic carrier concentration and can therefore influence the position of the Fermi level. The details of crystal growth were discussed in Chap. 17.

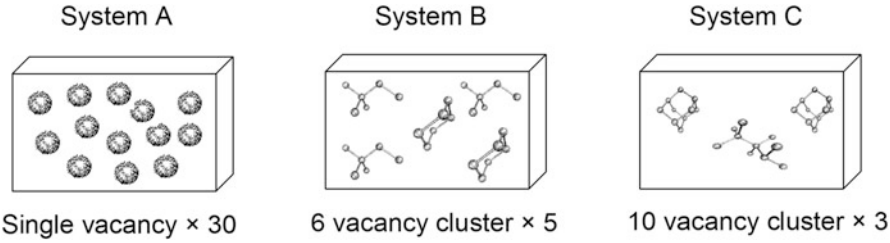
19.8 Summary

In this chapter, we discussed defects as imperfections that disturb the periodic structure of the crystal. The defects were classified into four groups according to their structural dimension. Point defects (0D), line defects (1D), planar defects (2D), and volume defects (3D) were explained. Several characterization techniques were introduced, and some issues regarding semiconductor single crystal growth were also discussed.

Problems

1. Give some examples of physical properties that defects can change.
2. Identify the types of point defects shown in Fig. 19.1. Please re-sketch the figure.
3. Calculate the number of vacancies per cubic meter in iron at 750 °C. The energy for vacancy formation is 1.08 eV/atom. Also, the density and atomic weight for Fe are 7.65 $g \cdot cm^{-3}$ and 55.85 $g \cdot mol^{-1}$, respectively. Assume A is unity.

4. Find the equilibrium concentration of defects for $T = 0$, 200, 400, 600, 800, 1000, and 1200 K if the energy to form a defect is 1 eV/atom. Assume A is unity. Graph your results. For T = 1200 K, how many atoms per single vacancy are present?
5. The formation energies of vacancy clusters in Si are listed below. Calculate the formation energy of (i) system a (30 single vacancies), (ii) system B (five 6-vacancy clusters), and (iii) system C (three 10-vacancy clusters). Which system has the lowest formation energy? Why?



(Cluster shapes reprinted with permission from Europhys Lett Vol. 43, p. 697, Fig. 1, Bongiorno A, Colombo L, and Diaz de la Rubia T, "Structural and binding properties of vacancy clusters in silicon," p. 697. Copyright 1998, EPD Sciences.)

Size	1	6	10
Energy (eV)	3.4	11.4	15.6

6. Briefly describe the difference between an edge dislocation and a screw dislocation.
7. Show how to find the Burgers' vector for a screw dislocation.
8. GaAs/InAs have a 7.2% lattice mismatch. How many monolayers of InAs may be grown on GaAs before a semi-coherent boundary is formed? ($a_{GaAs} = 0.565$ nm $a_{InAs} = 0.606$ nm, assume $b = a_{InAs}/\sqrt{2}$).
9. What is preferential etching?
10. What have been the goals of the semiconductor industry in silicon crystal growth technology? Why?

References

Callister WD (2000) Materials science and engineering: an introduction, 5th edn. John Wiley & Sons, Inc., New York

Wolfe CM, Holonyak N Jr, Stillman GE (1989) Physical properties of semiconductors. Prentice-Hall, Englewood Cliffs, NJ

Further Reading

Adachi S (1992) Physical properties of III-V semiconductor compounds: InP, InAs, GaAs, GaP, InGaAs, and InGaAsP. Wiley, New York, pp 263–286

Anderson JC, Leaver KD, Leevers P, Rawlings RD (2003) Materials Science for Engineers. Nelson Thornes Ltd, Cheltenham, UK

Bongiorno A, Colombo L, Diaz de la Rubia T (1998) Structural and binding properties of vacancy clusters in silicon. Europhys Lett 43:695–700

Hayes W, Stoneham AM (2004) Defects and defect processes in nonmetallic solids. Dover Publications, New York

Hurle DTJ, Rudolph P (2004) A brief history of defect formation, segregation, faceting, and twinning in melt-grown semiconductors. J Cryst Growth 264:550–564

Hurle DTJ (2004) Point defects in compound semiconductors. In: Muller G, Metois JJ, Rudolph P (eds) Crystal growth-from fundamentals to technology. Elsevier, Oxford, pp 323–343

Kittel C (1986) Introduction to solid state physics. John Wiley & Sons, New York

Murr LE (1978) Solid-state electronics. Marcel Dekker, New York

Nalawa HS, Bloembergen N, Laureate N (2000) Handbook of advanced electronic and photonic materials and devices. Academic Press, San Diego

Shaffner TJ (1997) Characterization challenges for the ULSI era. In: Rai-Choudhury P, Benton JL, Schroder DK, Shaffner TJ (eds) Proceedings of the Electrochemical Society symposium on diagnostic techniques for semiconductor materials and devices. Electrochemical Society, Pennington, NJ, pp 1–13

Swaminathan V, Macrander AT (1991) Materials aspects of GaAs and InP based structures. Prentice-Hall, Englewood Cliffs, NJ

Appendices

A.1 Physical Constants

Angstrom unit	Å	10^{-10} m = 10^{-8} cm = 10^{-4} μm
Avogadro constant	$\mathcal{N}_A$	6.02204×10^{23} mol^{-1}
Bohr radius	a_0	0.52917 Å
Boltzmann constant	k_b	1.38066×10^{-23} J·K^{-1} ($= R/\mathcal{N}_A$)
		8.61738×10^{-5} eV·K^{-1}
Calorie	cal	4.184 J
Elementary charge	q	1.60218×10^{-19} C
Electron rest mass	m_0	0.91095×10^{-30} kg
Electron Volt	eV	1.60218×10^{-19} J
		23.053 kcal·mol^{-1}
Gravitational constant	g	9.81 m·s^{-2}
Gas constant	R	1.98719 cal·mol^{-1}·K^{-1}
		8.31440 J·mol^{-1}·K^{-1}
Permeability in vacuum	μ_0	$4\pi 10^{-9} = 1.25633 \times 10^{-6}$ H·m^{-1}
Permittivity in vacuum	ε_0	8.85418×10^{-12} F·m^{-1} ($=1/\mu_0 c^2$)
Planck's constant	h	6.62617×10^{-34} J·s
Reduced Planck's constant	$\hbar$	1.05458×10^{-34} J·s ($= h/2\pi$)
Proton rest mass	M_p	1.67264×10^{-27} kg
Standard atmosphere	atm	1.01325×10^5 N·m^{-2}
Thermal voltage at 300 K	k_bT/q	0.0259 V
Velocity of light in vacuum	c	2.99792×10^8 m·s^{-1}
Wavelength of 1-eV quantum	λ	1.23977 μm

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7

A.2 International System of Units (SI Units)

Base units

Quantity	*Unit name*	*Unit symbol*
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

Prefixes

Factor	*Prefix*	*Symbol*	*Factor*	*Prefix*	*Symbol*
10^{24}	yotta	Y	10^{-1}	deci	d
10^{21}	zetta	Z	10^{-2}	centi	c
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	μ
10^{12}	tera	T	10^{-9}	nano	n
10^{9}	giga	G	10^{-12}	pico	p
10^{6}	mega	M	10^{-15}	femto	f
10^{3}	kilo	k	10^{-18}	atto	a
10^{2}	hecto	h	10^{-21}	zepto	z
10^{1}	deka	da	10^{-24}	yocto	y

Derived units

Quantity	*Special name*	*Unit symbol*	*Dimension*
Angle	radian	–	rad
Solid angle	steradian	–	sr
Speed, velocity	–	–	$m \cdot s^{-1}$
Acceleration	–	–	$m \cdot s^{-2}$
Angular velocity, frequency	–	$rad \cdot s^{-1}$	
Angular acceleration	–	–	$rad \cdot s^{-2}$
Frequency	hertz	Hz	s^{-1}
Force	newton	N	$kg \cdot m \cdot s^{-2}$
Pressure, stress	pascal	Pa	$N \cdot m^{-2}$
Work, energy, heat	joule	J	$N \cdot m$, $kg \cdot m^2 \cdot s^{-2}$
Power	watt	W	$J \cdot s$
Electric charge	coulomb	C	$A \cdot s$

Electric potential	volt	V	$J \cdot C^{-1}$, $W \cdot A^{-1}$
Resistance	ohm	Ω	$V \cdot A^{-1}$
Conductance	siemens	S	$A \cdot V^{-1}$, Ω^{-1}
Magnetic flux	weber	Wb	V·s
Inductance	henry	H	$Wb \cdot A^{-1}$
Capacitance	farad	F	$C \cdot V^{-1}$
Electric field strength	–	–	$V \cdot m^{-1}$, $N \cdot C^{-1}$
Magnetic induction	tesla	T	$Wb \cdot m^{-2}$, $N \cdot A^{-1} \cdot m^{-1}$
Electric displacement	–	–	$C \cdot m^{-2}$
Magnetic field strength	–	–	$A \cdot m^{-1}$
Celsius temperature	degrees Celsius	°C	K
Luminous flux	lumen	lm	cd·sr
Illuminance	lux	lx	$lm \cdot m^{-2}$
Radioactivity	becquerel	Bq	s^{-1}
Catalytic activity	katal	kat	$mol \cdot s^{-1}$

A.3 Physical Properties of Elements in the Periodic Table

The following figures summarize the general physical properties of most elements in the periodic table. These include their natural forms (Fig. A.1) with the structure in which they crystallize, their density of mass (Fig. A.2), boiling point (Fig. A.3), melting point (Fig. A.4), thermal conductivity (Fig. A.5), molar volume (Fig. A.6), specific heat (Fig. A.7), atomic radius (Fig. A.8), oxidation states (Fig. A.9), ionic radius (Fig. A.10), electronegativity (Fig. A.11), and electron affinity (Fig. A.12).

A.3.1 Chapter 2: Atomic Orbital

Generalities and Description

Since the discovery of the Schrödinger equation in 1925, it is well known that quantum particles such as electrons can be described as both particles (in the classical approach) and as *wave*. The wave description of an electron states implies that position cannot be fully known; instead, a wavefunction is required to describe the *probability* of finding any electron in a given region.

When electrons are linked to the nucleus of an atom, they cannot occupy all the available positions around the nucleus, and their position is quantified in *confined regions called atomic orbital*. An atomic orbital is a mathematical function describing the wave behavior of electrons in one of these confined regions (Fig. A.13).

Each atomic orbital is distinguished by three quantum numbers: n, l, and m.

n is a positive integer called the *principal quantum number*, l is an integer between 0 and $n - 1$ called the *azimuthal or angular quantum number*, and m is an integer between $-l$ and l called the *magnetic quantum number*.

Fig. A.1 Natural forms of elements in the periodic table

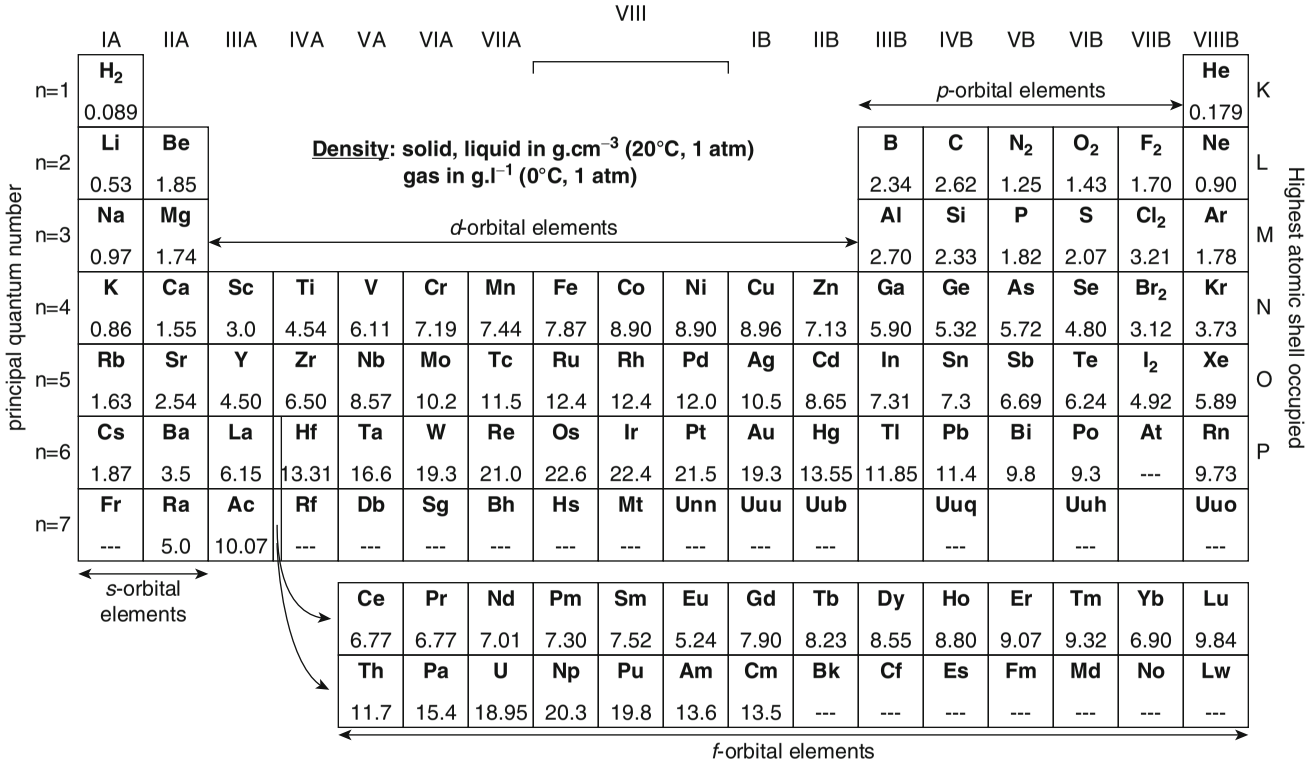


Fig. A.2 Density of mass of elements in the periodic table

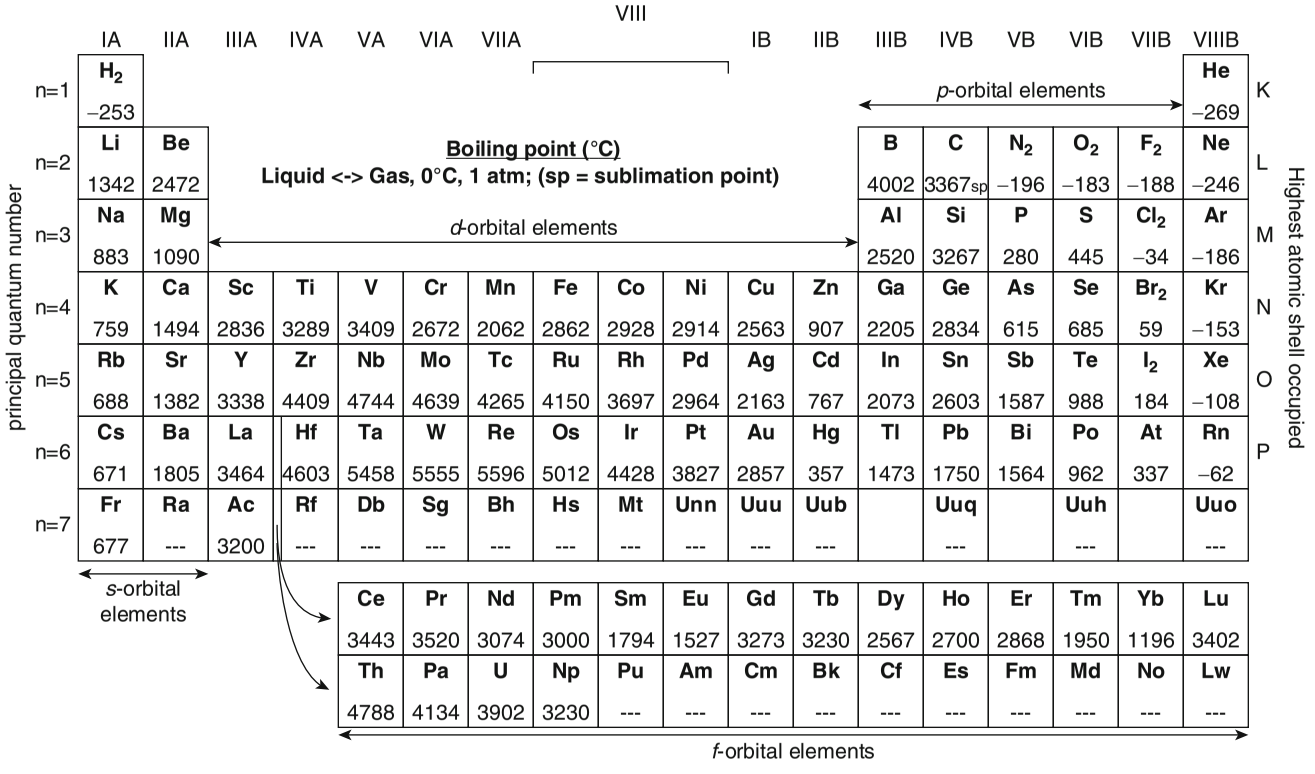


Fig. A.3 Boiling point of elements in the periodic table

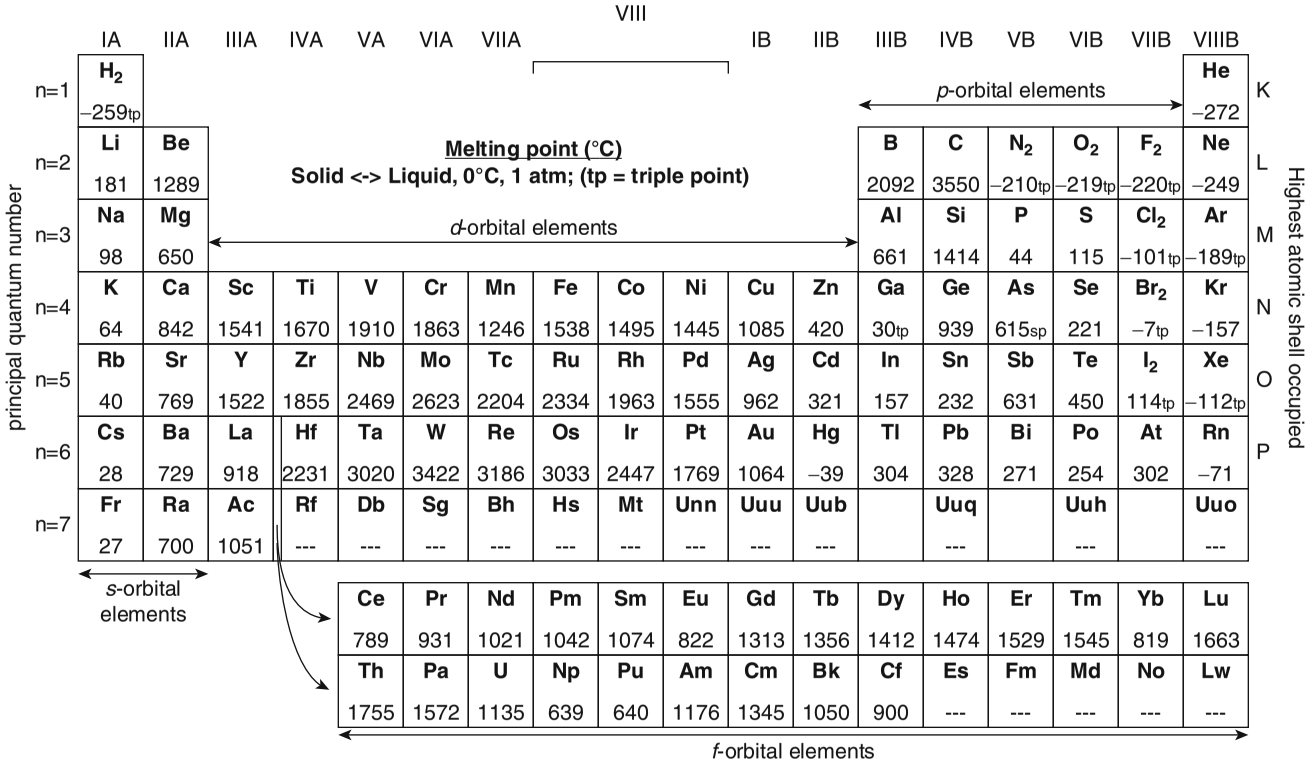


Fig. A.4 Melting point of elements in the periodic table

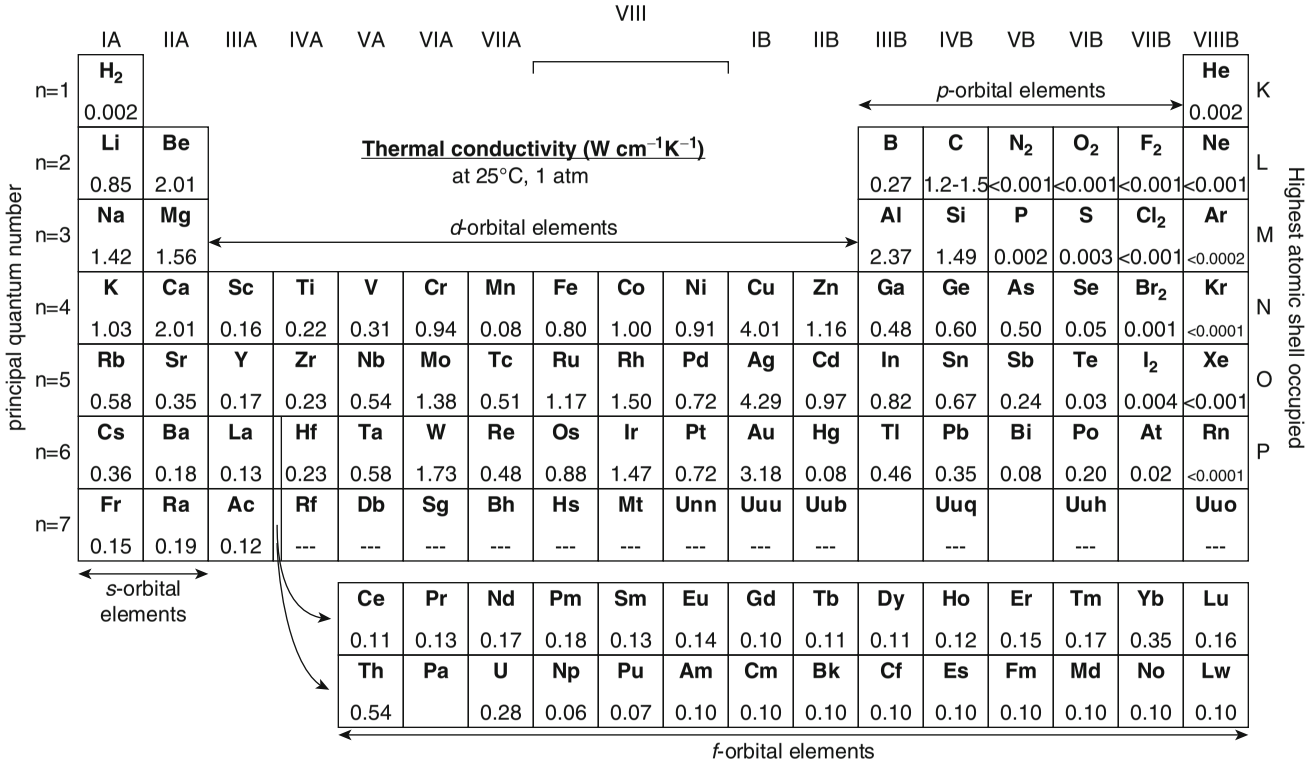


Fig. A.5 Thermal conductivity of elements in the periodic table

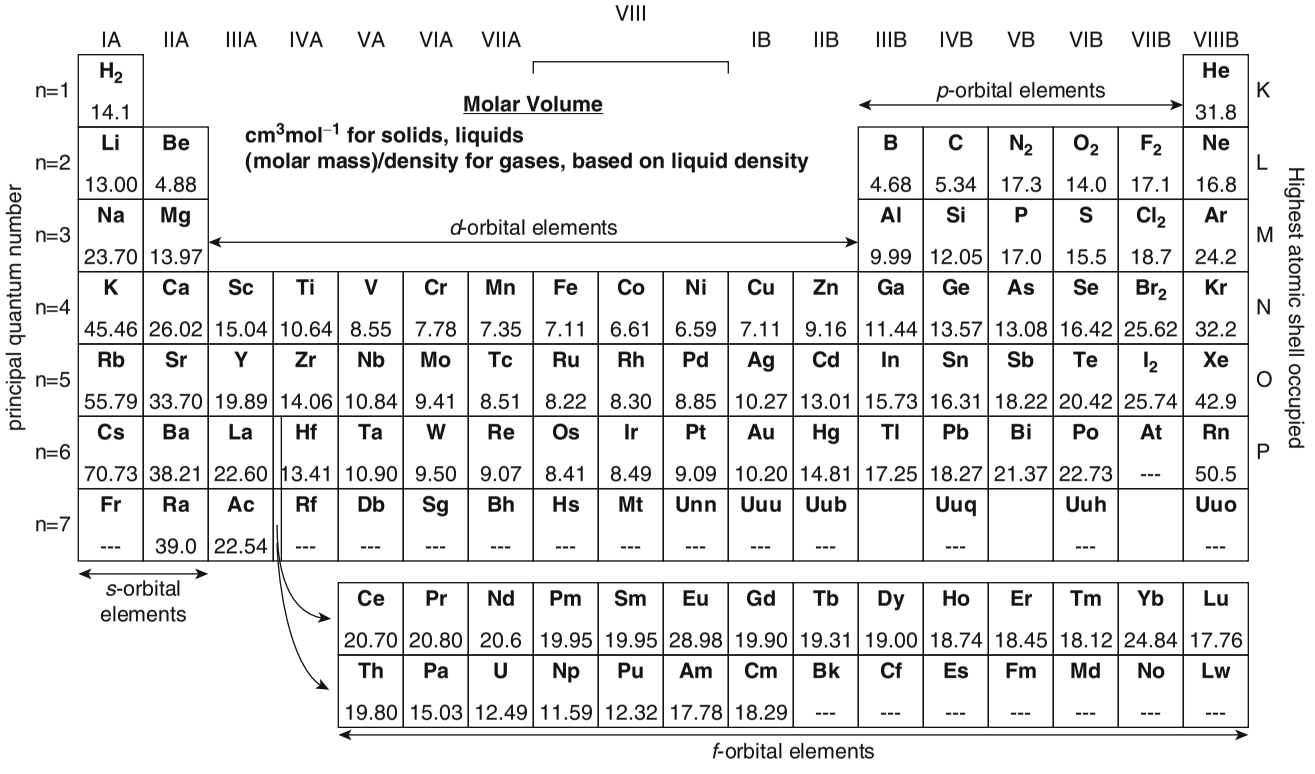


Fig. A.6 Molar volume of elements in the periodic table

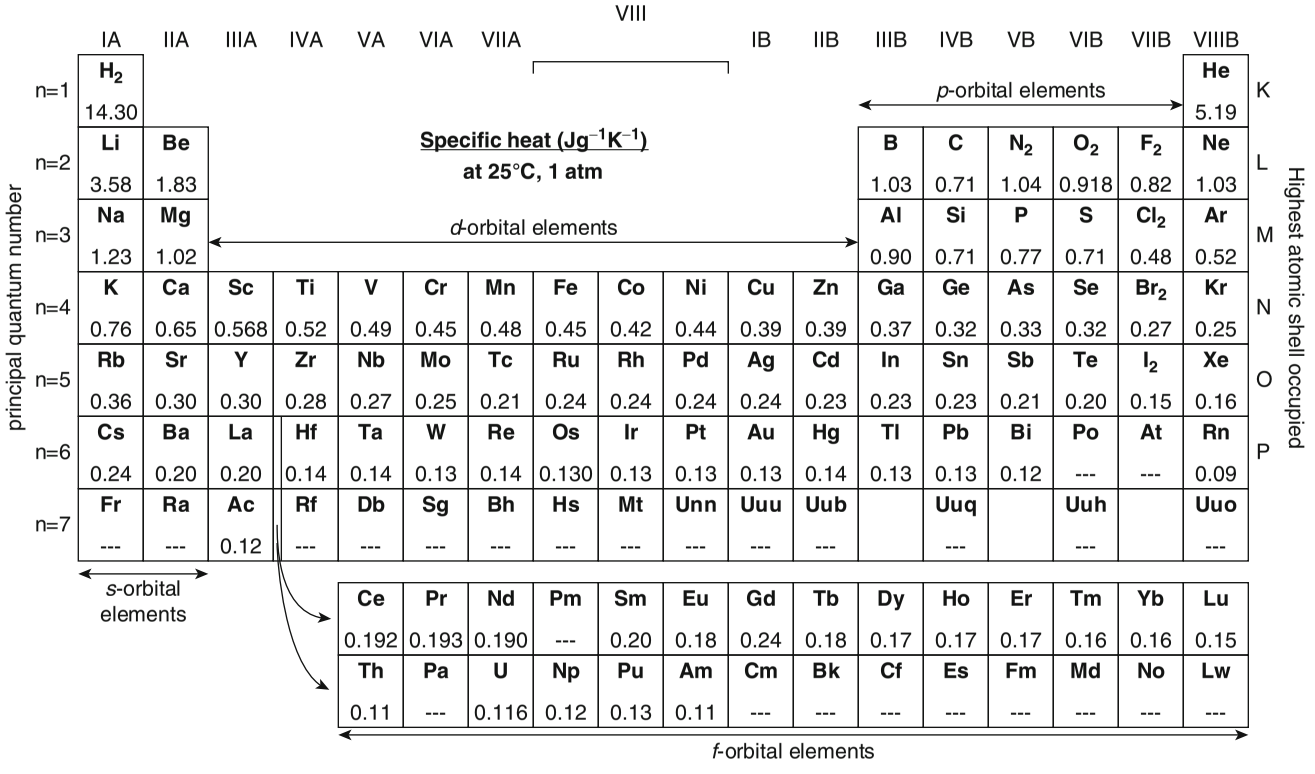


Fig. A.7 Specific heat of elements in the periodic table

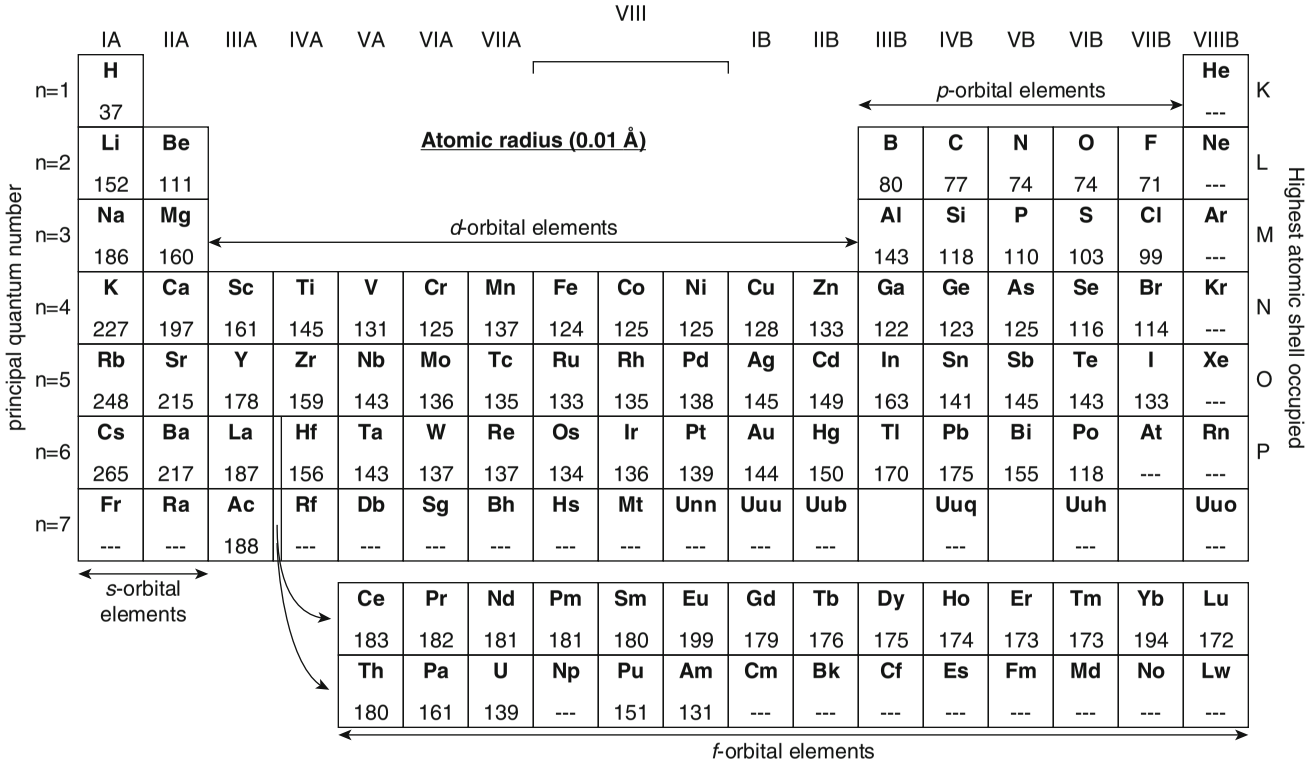


Fig. A.8 Atomic radius of elements in the periodic table

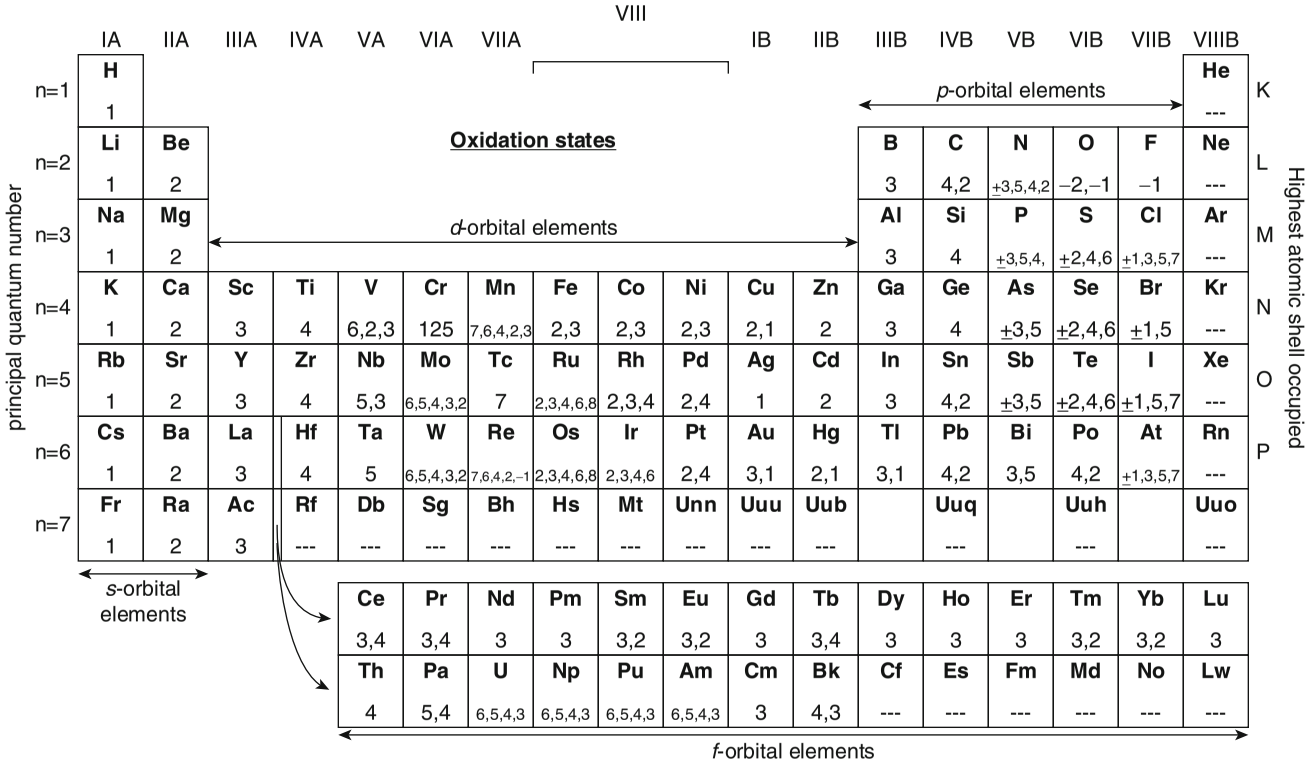


Fig. A.9 Oxidation states of elements in the periodic table

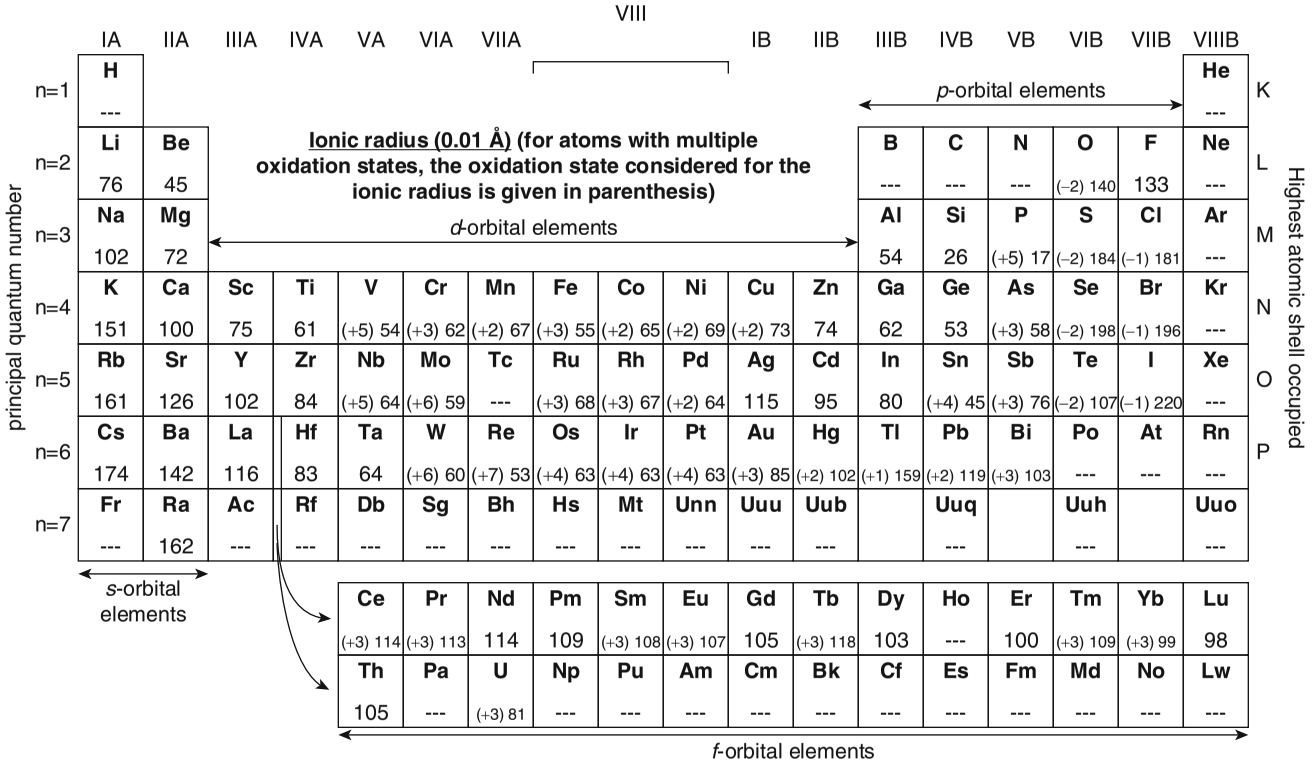


Fig. A.10 Ionic radius of elements in the periodic table

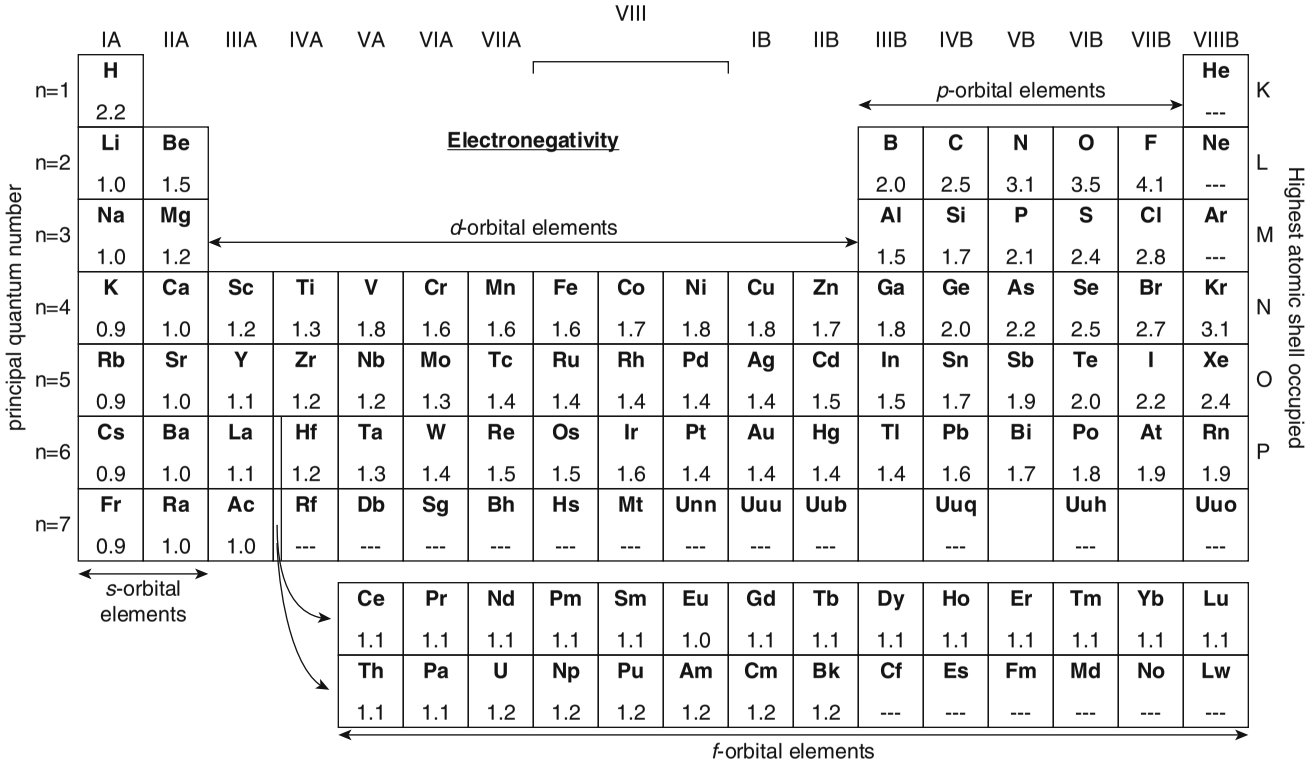


Fig. A.11 Electronegativity of elements in the periodic table

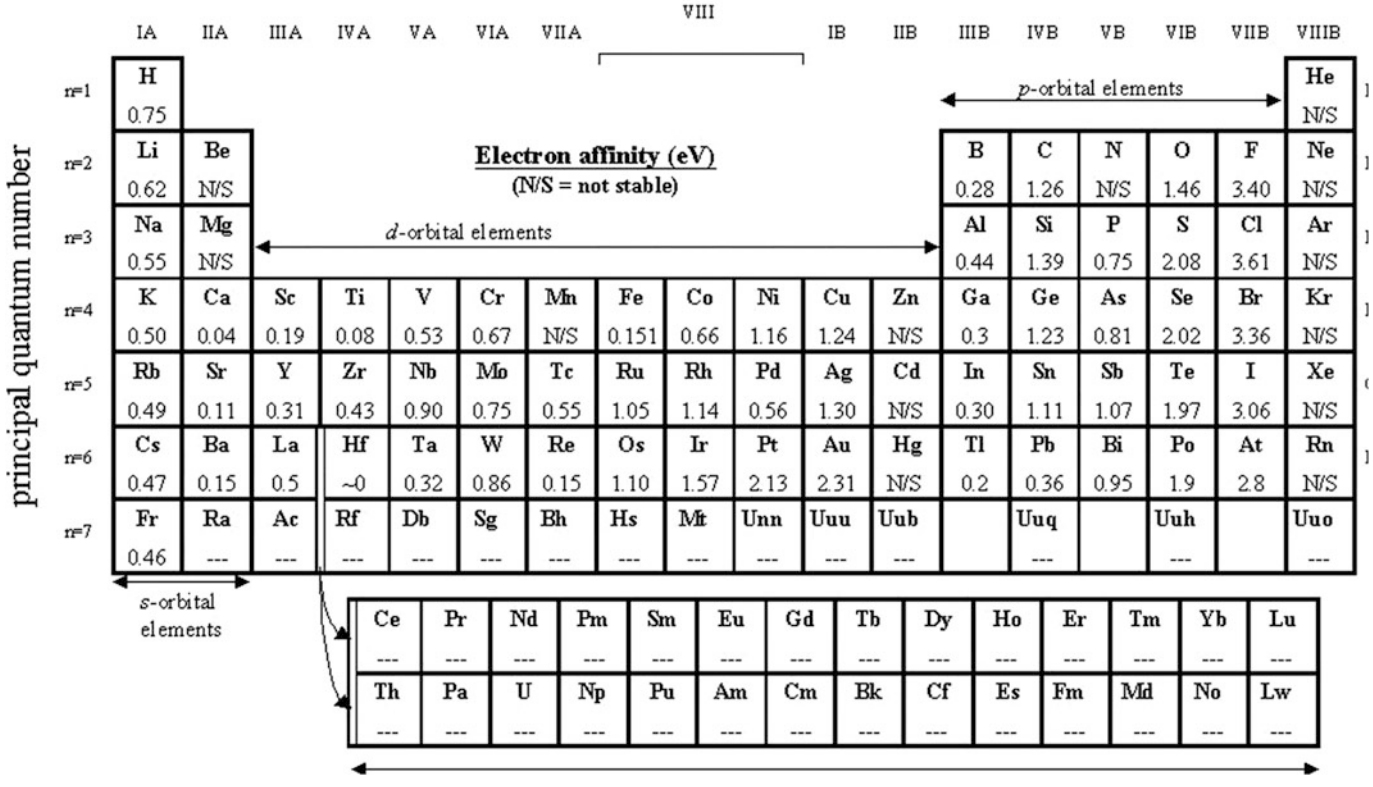


Fig. A.12 Electron affinity of elements in the periodic table

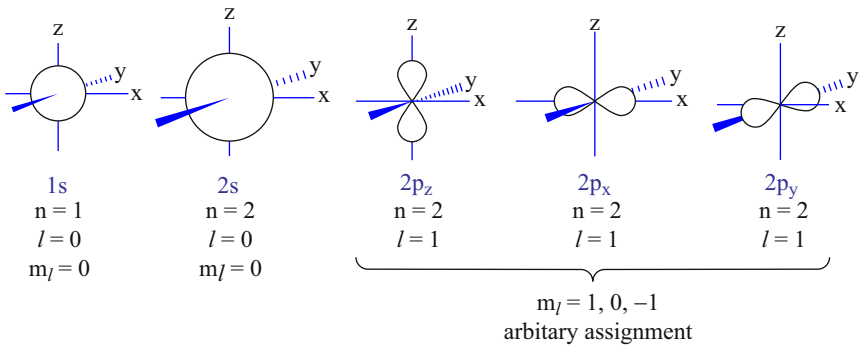


Fig. A.13 Sketch of 1s, 2s, and 2p atomic orbital wavefunctions

Any given (n, l, m) corresponds to one atomic orbital and can hold two electrons because of the Pauli principle. These two electrons thus have opposite spin.

The principal quantum number n refers to one shell of electrons around the nucleus and describes the size of orbitals. The higher this number, the farther is the shell from the nucleus. The completion of this layer with electrons describes the number of covalent bond which can be formed, and thus the n shell of an atom is primordial I the description of its properties.

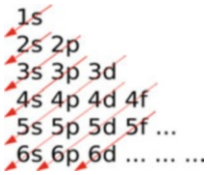
For a given n, the azimuthal number l refers to one particular orbital atomic in the n shell and describes the shape of the orbital.

l = 0 corresponds to an s orbital (sharp)
l = 1 corresponds to a p orbital (principal)
l = 2 corresponds to a d orbital (diffuse)
l = 3 corresponds to an f orbital (fundamental)

The magnetic quantum number m describes the orientation of orbitals in space.

For a given atom of known atomic number, the atomic electrons occupy orbitals from the lower shells with low-energy to the highest-energy shells.

The *Aufbau principle* (or Pauli rule) states which orbital are occupied by electrons for a given atom:



In this representation, the number corresponds to the principal quantum number n, while the letter (s, p, d, or f) gives the azimuthal quantum number l. The magnetic

quantum number is not represented and is suggested by the value of l since m is between −l and l. The red arrow gives the order in which the atomic orbital is filled. This order is the same for all atoms.

Example 1 Helium (He), two electrons.

First, the 1s orbital is filled. This orbital corresponds to $n = 1, l = 0$. Since m is between −l and l, the only value possible for m is 0. Thus the 1s orbital corresponds to $(n, l, m) = (1, 0, 0)$. The Pauli principle states that this orbital can hold two electrons. Thus for helium, the electronic configuration is $1s^2$.

Example 2 Carbon (C), 6 electrons.

First the 1s orbital is filled and can hold two electrons, similarly to the helium atom.

Then the 2s orbital is filled. This orbital corresponds to $n = 2, l = 0$, which forces $m = 0$. Thus, the 2s orbital corresponds to the $(n, l, m) = (2, 0, 0)$ configuration and can hold two electrons.

Then the 2p orbital is filled. This orbital corresponds to $n = 2$ and $l = 1$. Since $l = 1$, the values possible for m are -1, 0, and 1. Thus the 2p orbital contains the three configurations $(2, 1, -1)$, $(2, 1, 0)$, and $(2, 1, 1)$ and can hold a total of six electrons. The last two electrons of carbon can thus be contained in the 2p orbital. This 2p is not completely filled for the carbon atom and can hold four more electrons.

This explains why the carbon atom can be involved in four covalent bonds.

The electronic configuration of carbon is written $1s^2\ 2s^2\ 2p^2$ or [He] $2s^2\ 2p^2$.

A question remains unanswered: There are three p orbitals (p_x, p_y, p_z) in the 2p subshell, and these orbitals of same energy levels, called degenerate levels, are equivalent. Does the second electron go into the same orbital as the first, or does it go to another p orbital? To answer this question, we can use Hund's rule. This rule states that when filling the orbitals, one electron is added to each degenerate orbital before two electrons are added to the same orbital. Moreover, all the first electrons added to the degenerate orbitals have the same spin orientation.

As a result, the electrons in the 2p orbitals for carbon can be represented as follows:

A.3.2 Tight-Binding Method

For being one the most fundamental theoretical calculation tool in crystalline structure band energy modeling, multiple books concerning tight-binding method can be found in the literature, such as in Razeghi (1989). It is similar to the LCAO

(linear combination of atomic orbital) used in chemistry: Both models are approximation and rely on the periodicity of crystalline structures in order to model the interactions between atoms.

Understanding the tight-binding method is primordial, and this textbook assumes that its basic concepts are known. Yet, this appendix offers an overview of the method based on the example of a one-dimensional periodic structure composed of one type of atom. The similarity between the results obtained here and the equations found in the chapter can be easily noticed.

Bloch's Theorem

Let's consider a crystalline structure in which the atoms are assumed to be perfectly arranged, so that they introduce a periodic, infinite potential. This potential V has the same periodicity as the crystalline structure.

Without loss of generalities, the Schrödinger equation for an electron inside the crystalline structure can be written:

$$\widehat{\mathrm{H}}\psi\left(\vec{r}\right) = \left(-\frac{\hbar^2}{2m}\Delta + V\left(\vec{r}\right)\right)\psi(r) = E\psi\left(\vec{r}\right) \tag{1}$$

where E is the energy eigenvalue of the Hamiltonian operator.

The *Bloch's theorem* states that given a periodic potential V, the eigenstates ψ of the one-electron Hamiltonian (1) with periodic potential can be written as the product of a plane wave with a periodic function in the Bravais lattice of the solid. In other words:

$$\psi_{\vec{nk}}\left(\vec{r}\right) = e^{i\vec{k}\vec{r}}u_{\vec{nk}}\left(\vec{r}\right) \tag{2}$$

where $u_{\vec{nk}}\left(\vec{r}+\vec{R}\right) = u_{\vec{nk}}\left(\vec{r}+\vec{R}\right)$ for any $\vec{R}$ of the Bravais lattice.From Bloch's theorem (2), we get the following result: Any wavefunction which is eigenstate of a Hamiltonian with periodic potential verifies:

$$\psi_{\vec{nk}}\left(\vec{r}+\vec{R}\right) = e^{i\vec{k}\vec{R}}\psi_{\vec{nk}}\left(\vec{r}\right) \tag{3}$$

This means that when the physical space in real space is shifted of a vector from the Bravais lattice $\vec{R}$, only the phase of the wavefunction is affected.

The Tight-Binding Method: Generalities

Let's define the $\phi_n\left(\vec{r}\right)$ atomic orbitals, which are eigenstates of the Hamiltonian $\mathrm{H_{at}}$ of a single atom. When this atom is placed inside a crystalline structure, the electrons inside the structure can no longer be considered as a collection of isolated electrons, since the low distance between atoms makes valence electrons interact with each other. In other words, the atomic orbitals overlap adjacent atomic sites and thus are no longer eigenstates of the new Hamiltonian of the crystal. One of the assumptions

of the tight-binding model is that the Hamiltonian of the crystal can be still described as the sum of Hamiltonian of each single atom with a small perturbation:

$$\widehat{\mathrm{H}}\left(\vec{r}\right) = \sum_{\vec{R}_n} \widehat{\mathrm{H}}_{at}\left(\vec{r} - \vec{R}_n\right) + \Delta U\left(\vec{r}\right) \tag{4}$$

where $\vec{R}_n$ locates all the atomic sites inside the crystal and $\Delta U\left(\vec{r}\right)$ is the potential energy of interaction between atomic sites and is considered as a perturbation.

Since the effect of interaction between atoms is considered small, in the tight-binding method, one writes the solution of the Hamiltonian crystal as combination of atomic orbitals described above. In other words, we write any eigenstate of the Hamiltonian (3) in this form:

$$\psi_k\left(\vec{r}\right) = \sum_{\vec{R}_n} \sum_m c_m\left(\vec{R}_n\right)\phi_m\left(\vec{r} - \vec{R}_n\right) \tag{5}$$

where the first sum is over each atom of the crystalline structure and the second sum is over the different atomic orbitals of a single isolated atom. The $c_m\left(\vec{R}_n\right)$ are constants that need to be solved. The two following steps are dedicated to finding the value of these constants.

First of all, the Bloch's theorem states that $\psi_k\left(\vec{r}\right)$ has the same period as the crystalline structure, so that we can write:

$$\psi_k\left(\vec{r} + \vec{R}_n\right) = e^{i\vec{k}\vec{R}_n}\psi_k\left(\vec{r}\right) \tag{6}$$

The reader can easily verify that this equation leads to the following result:

$$b_m\left(\vec{R}_n\right) = e^{i\vec{k}\vec{R}_n} b_m\left(\vec{0}\right) \tag{7}$$

Then, the normalization of wavefunction to unity can be written:

$$\int \psi_k{}^*\left(\vec{r}\right)\psi_k\left(\vec{r}\right)d^3r = 1 \tag{8}$$

which after simplifications and neglecting atomic overlap integrals gives the following result:

$$b_m\left(\vec{0}\right) \simeq \frac{1}{\sqrt{N.}} \tag{9}$$

Using (4), (5), and (6) that the actual form of any eigenstate is:

$$\psi_k(\vec{r}) \simeq \frac{1}{\sqrt{N}} \sum_{\vec{R_n}} \sum_m e^{i\vec{k}\vec{R_n}} \phi_m(\vec{r} - \vec{R_n}) \tag{10}$$

In this equation, the first sum is labeled over the position $\vec{R_n}$ of every atom in the crystalline structure, and the second sum is labeled over the different valence orbitals of a given isolated atom.

Example

Let's give an example of use of the tight-binding method for a one-dimensional crystalline structure composed of one type of atom. Moreover, we assume that this atom only has one valence atomic orbital.

As seen in the previous section, any eigenstate of the Hamiltonian can be written $\psi_k(\vec{r}) \simeq \frac{1}{\sqrt{N}} \sum_{\vec{R_n}} \sum_m e^{i\vec{k}\vec{R_n}} \phi_m(\vec{r} - \vec{R_n})$. For this example, $m = 1$, since the atom only has one valence orbital. Let N be the total number of atoms in the structure, and let's use a more condensed formalism where $| k >= \psi_k(\vec{r})$ and $| n >= \phi(\vec{r} - \vec{R_n})$, so that:

$$| k >= \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{inka} | n >$$

where a is the distance between atoms.

We then assume that the orbital wavefunctions of only closest atoms overlap so that in the Hamiltonian matrix, all components $<i|H|j> = H_{i,j}$ are equal to zero, except the tridiagonal values.

In other words:

$$\begin{cases} < i|H|j >= 0 \quad if \ \ | i - j | > 1 \\ < i|H|i >= E_0 \\ < i \pm 1|H|i >= -\Delta \end{cases}$$

where $-\Delta$ can be interpreted as the bonding energy between atoms. In addition, we write the normalization equation between atoms wavefunction as:

$$< i|i >= 1 \quad \text{and} \quad < i \pm 1|i >= S < 1$$

Since $|k>$ is eigenstate of the Hamiltonian, we can determine its corresponding eigenvalue by writing:

$$\widehat{H}|k >= E|k >= \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{inka} \widehat{H} | n >$$

which, after multiplying on the left by $<k|$, leads to:

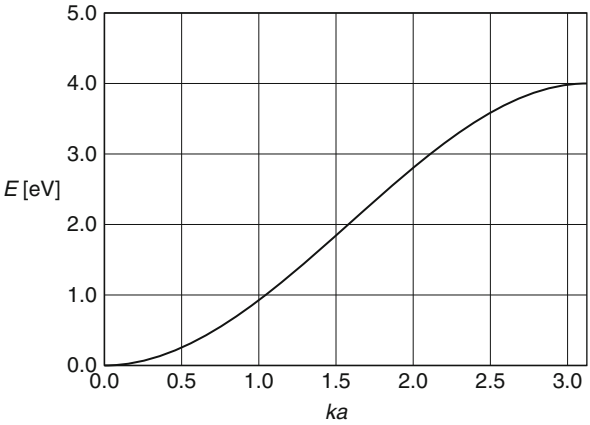
$$< k \mid \widehat{H}|k >=< k|E_k|k >= E_k = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{N} e^{i(n-m)\mathrm{ka}} < m|H|n >$$

And:

$$E_k = \frac{1}{N}\sum_{n=1}^{N} < n|H|n > + \frac{1}{N}\sum_{n=1}^{N} < n-1|H|n > e^{+\mathrm{ika}} + \frac{1}{N}\sum_{n=1}^{N} < n-1|H|n > e^{-\mathrm{ika}}$$

Thus the dispersion relation inside the crystal can is:

$$E_k = E(k) = E_0 - 2\Delta\cos(ka)$$



The result obtained here can be reproduced for different crystal structures and atomic orbitals.

A.4 Physical Properties of Important Semiconductors

Semiconductor		Bandgap energy (eV) 300 K	0 K	Band	ε
Element	C	5.47	5.48	Indirect	5.7
	Si	1.12	1.17	Indirect	11.9
	Ge	0.66	0.74	Indirect	16.0
	Sn		0.082	Direct	
IV-IV	α-SiC	2.996	3.03	Indirect	10.0

(continued)

Semiconductor		Bandgap energy (eV)		Band	
		300 K	0 K		ε
III-V	BN	~7.5		Indirect	7.1
	GaN	3.36	3.50	Direct	12.2
	GaP	2.26	2.34	Indirect	11.1
	BP	2.0			
	AlSb	1.58	1.68	Indirect	14.4
	GaAs	1.42	1.52	Direct	13.1
	InP	1.35	1.42	Direct	12.4
	GaSb	0.72	0.81	Direct	15.7
	InAs	0.36	0.42	Direct	14.6
	InSb	0.17	0.23	Direct	17.7
II-VI	ZnS	3.68	3.84	Direct	5.2
	ZnO	3.35	3.42	Direct	9.0
	CdS	2.42	2.56	Direct	5.4
	CdSe	1.70	1.85	Direct	10.0
	CdTe	1.56		Direct	10.2
IV-VI	PbS	0.41	0.286	Indirect	17.0
	PbTe	0.31	0.19	Indirect	30.0

Semiconductor	Intrinsic carrier concentration at 300 K (cm^{-3})
Ge	2.4×10^{13}
Si	1.45×10^{10}
GaAs	2.15×10^{6}

Semiconductor		Mobility at 300 K (cm^2/Vs)		Effective masses (in units of m_0)	
		Electrons	Holes	Electrons m_e	Holes m_h
Element	C	1800	1200	0.2	0.25
	Si	1500	450	0.98[a] 0.19[b]	0.16[c] 0.49[d]
	Ge	3900	1900	1.64[a] 0.082[b]	0.04[c] 0.28[d]
	Sn	1400	1200		
IV-IV	α-SiC	400	50	0.60	1.00
III-V	BN				
	GaN	380		0.19	0.60
	GaP	100	75	0.82	0.60
	BP				
	AlSb	200	420	0.12	0.98
	GaAs	8500	400	0.067	0.082[c] 0.45[d]
	InP	4600	150	0.077	0.64

(continued)

Semiconductor		Mobility at 300 K (cm^2/Vs)		Effective masses (in units of m_0)	
		Electrons	Holes	Electrons m_e	Holes m_h
	GaSb	5000	850	0.42	0.04[c] 0.4[d]
	InAs	33,000	460	0.023	0.40
	InSb	80,000	1250	0.0145	0.40
II-VI	ZnS	165	5	0.40	
	ZnO	200	180	0.27	
	CdS	340	50	0.21	0.80
	CdSe	800		0.13	0.45
	CdTe	1050	100		
IV-VI	PbS	600	700	0.25	0.25
	PbTe	6000	4000	0.17	0.20

[a]Longitudinal effective mass
[b]Transverse effective mass
[c]Light-hole effective mass
[d]Heavy-hole effective mass

A.5 The Taylor Expansion

The Taylor expansion is a powerful mathematical method which yields a simple polynomial approximation for any mathematical function near a given point.

Let us consider a function f which can be differentiated at least $(n + 1)$ times at $x = x_0$. The Taylor expansion is such that the value of f at any point x can be determined from its value and that of its n consecutive derivatives at x_0 through:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots \\ \ldots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + R_n \tag{A.1}$$

where R_n is called the remainder and is equal to:

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1} \tag{A.2}$$

for an appropriate value ξ such that $|\xi - x_0| \leq |x - x_0|$.

As a result of this expansion, an approximate value of the function f near the point $x = x_0$ is obtained by neglecting the remainder R_n in Eq. (A.2). In principle, the more terms one chooses to keep in the expansion, the more accurate result one will get. R_n is used to evaluate the magnitude of the calculation error. It is often useful to carry the Taylor expansion near an extremum of the function f because some of its derivatives are then equal to zero, and a simplified expression is obtained.

A few examples of Taylor expansion for commonly used functions are given below:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{A.3}$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + - \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} \tag{A.4}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^3}{3!} - \frac{x^5}{5!} + - \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} \tag{A.5}$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^{n+1} x^n}{n} \tag{A.6}$$

There exist convergence ranges in evaluating the infinite sums inEq. (A.6) to Eq. (A.5). This means that the Taylor expansion will no longer be valid when trying to evaluate the sums for a value of x outside the convergence range. For example, the convergence range for e^x, $\sin(x)$, and $\cos(x)$ is $(-\infty, +\infty)$, whereas the convergence range for $Ln(1 - x)$ is $(-\infty, 1]$.

A.6 Fourier Series and the Fourier Transform

Fourier Series

A function $f(t)$ is periodic with a period T when it satisfies $f(t + T) = f(t)$ for any value of t. If such a periodic function is also piecewise continuous, then it can be written as the sum of trigonometric functions such that:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(nwt) + b_n \sin(nwt)\right) \tag{A.7}$$

where we have denoted $w = \frac{2\pi}{T}$, and:

$$a_0 = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)dt \tag{A.8}$$

$$a_n = \frac{2}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) \cos(nwt)dt \tag{A.9}$$

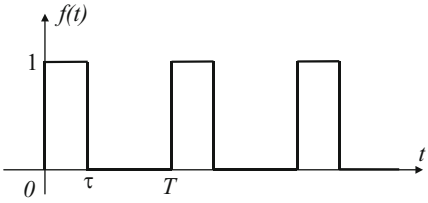


Fig. A.14 Example of periodic function used to illustrate the concept of Fourier series

$$b_n = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\sin(nwt)dt \tag{A.10}$$

Such a sum of trigonometric functions is called the Fourier series of $f(t)$, and the coefficients a_n and b_n are called its Fourier coefficients. The usefulness of such a mathematical expansion lies in its physical interpretation. Indeed, one can see that a periodic function of time can be decomposed into individual sine-like and cosine-like components, each periodic with a frequency nw where n is an integer. The magnitude of each component is given by the Fourier coefficients a_n and b_n. One can therefore obtain a "spectrum of frequencies" for the original function, which finds a number of applications in physics phenomena.

For example, the Fourier expansion of the function shown in Fig. A.14 is:

$$f(t) = \frac{\tau}{T} + \sum_{n=1}^{\infty}\frac{1}{n\pi}[\sin nw\tau\cos nwt + (1 - \cos nw\tau)\sin nwt] \tag{A.11}$$

Fourier Transformation

The Fourier transformation is a mathematical operation which consists of associating to a given function f a second function, called its Fourier transform F. The functions f and F do not operate on the same variables. The Fourier transform is similar to a Fourier series but can be applied to a general function $f(t)$ as long as it is pulse-like and $\int_{-\infty}^{\infty}|f(t)|dt < \infty$. Its Fourier transform F is then defined by:

$$F(w) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{-iwt}dt \tag{A.12}$$

Note that the Fourier transform F operates on frequencies w, whereas the original function f operates on time t. The Fourier transform plays the same role as the Fourier coefficients in Eq. (A.12), except that the summations on frequencies are now continuous rather than discrete. The original function f can be expressed in terms of its Fourier transform F through:

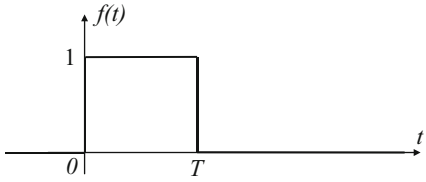


Fig. A.15 Example of an arbitrarily chosen function used to illustrate the concept of Fourier transform

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(w) e^{iwt} dw \tag{A.13}$$

For example, the Fourier transform of the function shown in Fig. A.15 is:

$$F(w) = \frac{1 - e^{-iwT}}{iw\sqrt{2\pi}} \tag{A.14}$$

A.7 The Pseudopotential Approach

When we want to calculate the band structure of a solid from first principles and write down the exact Hamiltonian of the system, we are confronted with a very difficult problem because not only do we have the Coulomb potential of the nuclear charges, but we also have the electron-electron interaction of the other electrons in the system to deal with. The way to avoid it is to make some simplifications, which keep the essence of the problem and make the solution tractable. We use the insight that we have and argue that, surely, it is possible to assume that the strongly bound full shells around the atom are not participating in the banding of the solid and they can be separated out, i.e., excluded from the banding electrons. The valence electrons can be treated separately and do not see the full potential of the nucleus. We know already from Chap. 4 that the outer shell electrons see a screened potential because the core electrons screen out the full nuclear attraction. But this is not all, we do not want to just take into account the screening, which is a many body effect, but go further and not allow the valence states to be mixed in the core states at all. So there are two effects to be considered. One is the screening, which can be considered to give rise to an effective nuclear charge and can be treated using the self-consistent "Hartree-Fock method". The other is projecting out the core eigenstates out of the solutions altogether. The latter is the pseudopotential method. In the pseudopotential method, one first decides which core states must be projected out. One does this by making the sought after Bloch wavefunctions orthogonal to these core states. Then one derives the effective potential for which these new Bloch states are the eigenfunction solution of the Schrödinger equation.

This procedure then makes the envelope wavefunction of the Bloch wave $u_{\vec{k}}$ in:

$$\Psi_{\vec{k}}(\vec{r}) = u_{\vec{k}}(\vec{r})e^{i\vec{k}.\vec{r}} \tag{A.15}$$

a much more smooth function than it would be if it were subject to the full or even screened Coulomb potential of the lattice ions. How do we find an approximation for this effective potential? In a naïve way, we have done this already in the Kronig-Penney model in Chap. 5. The Kronig-Penney model is indeed a truncated pseudopotential approximation to the true potential, but it is constructed in an ad hoc manner, without a well-defined prescription. The pseudopotentials used to calculate the band structure of solids are however derived using well-defined prescriptions.

One of the assumptions is that the basis states of all the electrons in the solid are constituted by core electron wavefunctions ϕ_j and valence electron wavefunctions χ_k and that these can be made orthogonal to each other. One then constructs the eigenstates of interest, namely, for the higher valence energy level states. These are built to avoid the core regions occupied by the core electrons. An example is as follows. Assuming $b_{n\vec{k}} = \sum_{\vec{s}} e^{i\vec{k}.s} b_n(\vec{r} - \vec{s})$ is a core function solution of the Schrödinger equation with energy E_n, we construct a more extended valence state which is made orthogonal to the core states and is of the form:

$$\Psi_{\vec{k}} = \sum_{\vec{g}} \alpha_{\vec{k}-\vec{g}} \chi_{\vec{k}-\vec{g}} \tag{A.16}$$

$$\chi_{\vec{k}} = e^{i\vec{k}.\vec{r}} - \sum_j a_j b_{j\vec{k}} \tag{A.17}$$

The a_j are selected to make the valence wavefunctions orthogonal to the core states. This new wavefunction has the core states projected out of it and is forced to also satisfy the Schrödinger equation. The projected states however introduce a new term in the SE which plays the role of a potential. The new term due to the core states, when combined with the old, gives us now an effective potential, which repels the valence electrons out of the core region, making the effective potential much more smooth than the original one. The pseudopotential method is a way of projecting out the core functions, out of what would normally be the total wavefunction, so that the more loosely bound valence functions avoid the region, which is normally filled by the core states. They do not see the strong Coulomb field anymore because they are forced to adopt a higher orbital or what is in effect a more loosely bound character near the core.

The two methods, Hartree-Fock self-consistent field or "HFT" method, which takes care of the potential of the other electrons and the Pauli principle, and the pseudopotential method, which forces the higher levels to avoid being core-like, can in principle be combined to produce an accurate band structure calculation. The HFT

method assumes that the Coulomb potential of the other electrons can be treated as an average potential, which can be evaluated self-consistently. It also assumes that the many particle wavefunctions are Slater determinants of Bloch functions so that they automatically satisfy the Pauli principle. The details of the "Pseudopotential and HFT" are beyond the scope of this book, and the reader is referred to the specialized works in the books by Ziman (1998), Callaway (1964), and Harrison (1966).

Further Reading

Callaway J (1964) Energy Band Theory. Academic Press, New York
Chuang SL (1995) Physics of Optoelectronic Devices. Wiley, New York
Harrison WA (1966) Pseudopotentials in the Theory of Metals. W.A. Benjamin, New York
Ziman JM (1998) Principles of the Theory of Solids. Cambridge University Press, Cambridge

A.8 The Monte-Carlo Method

Scattering in a Crystal

Electrons in a crystal with a given band structure can be considered as a collection of free particles. In the six-dimensional phase space of momentum $\vec{k}$ and space $\vec{r}$, we can represent each electron by a point of coordinates $(\vec{r}, \vec{k})$. As we have seen in Sect. 5.2.6, the motion of the electron is described by:

$$\hbar \frac{d\vec{k}}{dt} = q\left(\vec{E} + \vec{v} \times \vec{B}\right) \tag{A.18}$$

where $\vec{k}$ is the wavevector of the electron, q the electric charge, $\vec{E}$ the electric field, $\vec{v}$ the velocity, and $\vec{B}$ the magnetic field. In this appendix, we are going to study only the action of an electric field on the electron, so we put $\vec{B}=\vec{0}$. Under these conditions, the electrons start their journey by following a ballistic trajectory (they are freely accelerated). However, this motion is interrupted by collisions with atoms, impurities, etc., which we will consider as scattering events. As a result, the movement of the particles is far more complex, and it is useful to describe the motion of the electrons by a distribution function $f(\vec{k}, \vec{r}, t)$, which is the average occupancy of a point in the above phase space.

The time evolution of this function is described by the Boltzmann equation:

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \vec{\nabla}_{\vec{r}} f + \frac{d\vec{k}}{dt} \cdot \vec{\nabla}_{\vec{k}} f = \left(\frac{\partial f}{\partial t}\right)_{coll} \tag{A.19}$$

where, together with Eq. (A.18), the LHS describes all the ways the function evolves in phase space when subject to an electric and magnetic field.

If $\left(\frac{\partial f}{\partial t}\right)_{coll} dt$ describes the variation of the distribution during dt due to the collisions, the global variation can be written:

$$f\left(\vec{r} + d\vec{r}, \vec{k} + d\vec{k}, t + dt\right) = f\left(\vec{r}, \vec{k}, t\right) + \left(\frac{\partial f}{\partial t}\right)_{coll} dt \tag{A.20}$$

to first order:

$$f\left(\vec{r}, \vec{k}, t\right) + \frac{\partial f}{\partial t} dt + \vec{\nabla}_{\vec{r}} f \cdot d\vec{r} + \vec{\nabla}_{\vec{k}} f \cdot d\vec{k} = f\left(\vec{r}, \vec{k}, t\right) + \left(\frac{\partial f}{\partial t}\right)_{coll} dt \tag{A.21}$$

$$\frac{\partial f}{\partial t} + \vec{\nabla}_{\vec{r}} f \cdot \frac{d\vec{r}}{dt} + \vec{\nabla}_{\vec{k}} f \cdot \frac{d\vec{k}}{dt} = \left(\frac{\partial f}{\partial t}\right)_{coll} \tag{A.22}$$

Using $\vec{F} = \hbar \frac{d\vec{k}}{dt} = q\left(\vec{E} + \vec{v} \times \vec{B}\right)$, we get:

$$\frac{\partial f}{\partial t} + \vec{\nabla}_{\vec{r}} f \cdot \vec{v} + \vec{\nabla}_{\vec{k}} f \cdot \frac{\vec{F}}{\hbar} = \left(\frac{\partial f}{\partial t}\right)_{coll} \tag{A.23}$$

This equation states that the changes of the distribution function with time (represented by the first term on the LHS of this equation) are determined by the flow of electrons in real space (the second term in the LHS of the equation) and by the flow of electrons in $\vec{k}$-space (the last term in the LHS of the equation) and the collisions (right-hand side of the equation). The right-hand side describes the effects of the many different types of scattering mechanisms, which are active, including optical phonon scattering, acoustic scattering, impurity scattering, etc., so that it is often very difficult to solve for $f\left(\vec{k}, \vec{r}, t\right)$ analytically. However, given the scattering rates, a numerical solution or simulation of this equation, which is called the Monte-Carlo simulation, is always possible. This so-called Monte-Carlo method is a powerful tool and is becoming more and more popular.

Monte-Carlo Simulation

The idea of this method, introduced in the 1960s (see Shur 1990), is to simulate the motion of the particle in $\vec{k}$-space while keeping track of it in real space. In this model,

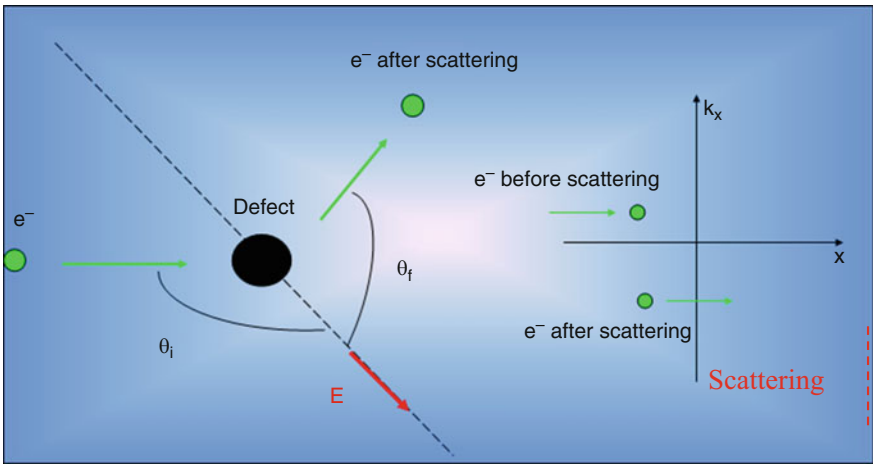


Fig. A.16 On the left, sketch of the scattering of an electron by an impurity. On the right, illustration of the disappearance and the appearance of the electron in the phase space

we consider that the motion of the electron is well described by Eq. (A.18) between two scattering events. But this free flight is interrupted by scattering processes that occur with a rate λ_i (i stands for the scattering process that we are considering). These processes are instantaneous events and change only the wavevector of the electron. They can be visualized as the particle disappearing and reappearing instantaneously at a different point of phase space (see Fig. A.16). If we observe a single electron for a sufficiently long time, the distribution of the times that the electron spends in the vicinity of different points in $\vec{k}$-space will reproduce the shape of $f\left(\vec{k}, \vec{r}, t\right)$.

The Monte-Carlo simulation can be divided into three different parts. First, we generate randomly with a computer the time remaining before the next scattering event. Then, between the two scattering events, we determine the motion of the electron using Eq. (A.22). Finally, we generate randomly the new direction of the wavevector.

- For the purpose of this simulation, we introduce a scattering rate $\Gamma = \sum_{i=1}^{n} \lambda_i\left(\vec{k}\right) + \lambda_0$ where we have introduced an artificial scattering mechanism, with a rate λ_0, so that Γ is a constant (finite). This self-scattering process interrupts the motion but does not change the momentum in any way. It can be described by the probability $W_0\left(\vec{k}, \vec{k}'\right) = \lambda_\circ\left(\vec{k}\right)\delta\left(\vec{k} - \vec{k}'\right)$. This rate is simply a mathematical tool used to make the global rate of the scattering events constant. In order not to change the rate too much, we choose λ_0 as small as possible. Thus, the probability of a scattering event between t and $t + dt$ can be described by $P(t)dt = e^{-\Gamma t}dt$. We

use this distribution of probabilities to generate random times t_s between one collision and the following one ($t_s = -1/\Gamma \ \ln(1-r)$, with r a random number between 0 and 1, follows this distribution).

- During these times t_s, the motion of the electron is well described by Eq. (A.18) with $\vec{B}=\vec{0}$ so that $\vec{k}(t) = \vec{k}_0 + \frac{q\vec{E}}{\hbar}t$, where $\vec{k}_0$ is the wavevector just after the previous collision, and $\vec{r}(t) - \vec{r}_0 = \int_{t_0}^{t} \vec{v}_g(t')dt' = \int_{t_0}^{t} \frac{1}{\hbar}\nabla_{\vec{k}} E dt'$, where $\vec{r}_0$ is the position of the particle in real space after the previous collision.
- The next step is to generate randomly the wavevector after each scattering event. But, before that, we need to determine which mechanism is responsible for the scattering. In order to find out which law we have to apply to generate the new wavevector, we assume that the probability of occurrence of one given process is proportional to its rate. To choose a mechanism, we generate randomly a number A, distributed with equal probability between 0 and Γ, and we test the inequality $\sum_{i=0}^{m} \lambda_i(\vec{k}) > A$. The first value of m satisfying this inequality is the scattering process we are going to use. We use the distribution function of probabilities of this mechanism to generate randomly the wavevector after the scattering event.

We repeat these three steps as long as we need to get a good approximation of $f(\vec{k}, \vec{r}, t)$. A criterion to stop our stimulation is to repeat the scenario until the differences in the drift velocity, for example, converge to a small enough number.

Thanks to this procedure, we are able to simulate the movement of the electron in the crystal. Then, we represent in a histogram the time that the electron spent in each cell of the phase space. It has been demonstrated that this histogram is proportional to the distribution function $f(\vec{k}, \vec{r}, t)$ when t tends to infinity.

Applications

The Monte-Carlo simulation is a useful tool to calculate quantities like the time spent in the valleys of a semiconductor or the diffusion coefficients of a material. It shows a good agreement with experiment as you can see in Fig. A.17. It can also be used to investigate the electron transport in small semiconductor devices. But this method only allows us to study a relatively small number of free electrons in the semiconductor: Typically 1 million electrons. The idea is that, for example, 1 million is enough to reproduce the behavior of all the particles. An example of a real space trajectory is shown in Fig. A.18.

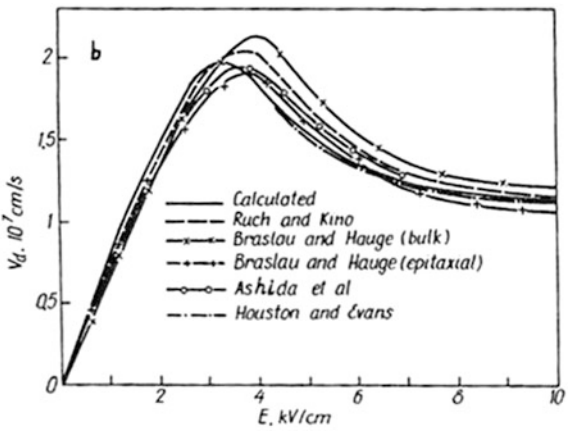


Fig. A.17 Measured and calculated drift velocity (Reprinted from Solid State Electronics Vol. 23, Pozhela, J., & Reklaitis, A., "Electron transport properties in GaAs at high electric fields," Fig. 7, p. 931, Copyright 2005, with permission from Elsevier)

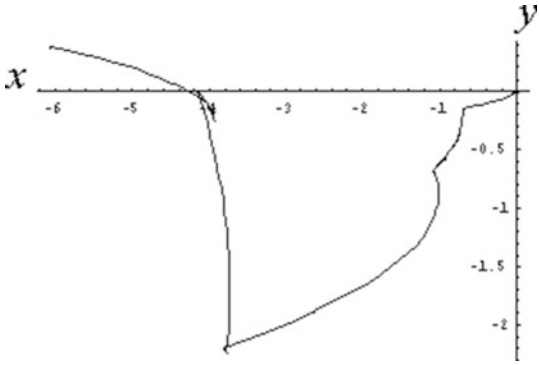


Fig. A.18 Simulation of the motion of an electron under an electric field E in the x-direction (ten collisions are simulated). The motion of the electron starts at the origin and evolves randomly. This figure represents the trajectory of the electron in real space

References

Pozhela J, Reklaitis A (1980) Electron transport properties in GaAs at high electric fields. Solid States Electron 23:927–933

Further Reading

Shur M (1990) Physics of Semiconductor Devices. Prentice-Hall, Englewoods Cliff

A.9 The Thermionic Emission

The thermionic emission theory is a semiclassical approach developed by Bethe (1942), which accurately describes the transport of electrons through a semiconductor-metal junction. The parameters taken into account are the temperature T, the energy barrier height $q\Phi_B$, and the bias voltage V between the far ends of the semiconductor and the metal. These quantities are illustrated in Fig. A.19.

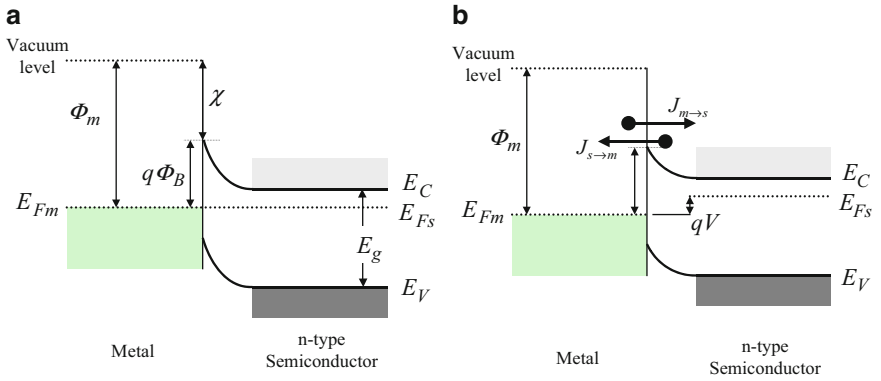


Fig. A.19 Energy band diagram of a Schottky metal-(*n*-type) semiconductor junction: (**a**) At equilibrium and (**b**) under forward bias ($V > 0$), showing the transport of electrons over the potential barrier as the main transport process under forward bias

The theory is based on the following three assumptions: (i) The energy barrier height $q\Phi_B$ at the interface is much higher than k_bT, (ii) the junction plane is at thermal equilibrium, and (iii) this equilibrium is not affected by the presence of an electrical current. By assuming these, the thermionic emission current only depends on the energy barrier height and not its spatial profile. Furthermore, the total current is therefore the sum of the current from the semiconductor into the metal, denoted $J_{s\to m}$, and that of the metal into the semiconductor, denoted $J_{m\to s}$.

To calculate the first current, $J_{s\to m}$, the theory assumes that the energy of the electrons in the conduction band is purely kinetic and that their velocity is distributed isotropically. The current density from the semiconductor into the metal can be calculated by summing the current contribution from all the electrons that have an energy higher than the barrier $q\Phi_B$ and that have a velocity component from the semiconductor toward the metal. This results in the following expression:

$$J_{s\to m} = \left(\frac{4\pi q m^* k_b^2}{h^3}\right) T^2 e^{-\frac{q\Phi_B}{k_bT}} e^{\frac{qV}{k_bT}} \tag{A.24}$$

or:

$$J_{s\to m} = A^* T^2 e^{-\frac{q\Phi_B}{k_bT}} e^{\frac{qV}{k_bT}} \tag{A.25}$$

where k_b is the Boltzmann constant, V is the bias voltage, Φ_B is the barrier height, T is the temperature in degrees Kelvin, h is Planck's constant, and m^* is the electron effective mass in the direction perpendicular to the junction plane, and $A^* = 4\pi q m^* k_b^2/h^3$ is called the effective Richardson constant for thermionic emission. This quantity can be related to the Richardson constant for free electrons, $A = 120\ \mathrm{A{\cdot}cm^{-2}{\cdot}K^{-2}}$, as discussed below.

Table A.1 Examples of values for A^*/A in a few semiconductors (Sze 1981)

Semiconductor	Si	Ge	GaAs
n-type <111>	2.2	1.11	0.068 (low field) 1.2 (high field)
n-type <100>	2.1	1.19	0.068 (low field) 1.2 (high field)
p-type	0.66	0.34	0.62

For *n*-type semiconductors with an isotropic electron effective mass m^* in the minimum of the conduction band, we have $A^*/A = m^*/m_0$, where m_0 is the electron rest mass.

For *n*-type semiconductors with a multiple-valley conduction band, the effective Richardson constant A^* associated with each local energy minimum is given by $A^*/A = \left(l_x^2 m_y^* m_z^* + l_y^2 m_z^* m_x^* + l_z^2 m_x^* m_y^*\right)^{1/2}/m_0$, where l_x, l_y, and l_z are the direction cosines corresponding to this energy minimum in the first Brillouin zone.

In the case of a *p*-type semiconductor, we need to consider the heavy-hole and the light-hole bands in the valence band, both of which have their maximum at the center of the Brillouin zone. The effective Richardson constant is then given by the following expression $A^*/A = \left(m_{lh}^* + m_{hh}^*\right)/m_0$, where m_{hh}^* and m_{lh}^* are the heavy-hole and light-hole effective masses, respectively. A few examples of values for A^*/A are given in Table A.1.

The second current contribution to the thermionic emission current is the current flowing from the metal into the semiconductor, $J_{m \to s}$. As the barrier height for the transport of electrons in this direction is independent of the applied bias voltage V (Fig. A.19b), $J_{m \to s}$ is also independent of the bias voltage. $J_{m \to s}$ is therefore equal to the opposite of $J_{s \to m}$ when $V = 0$, because no net current exists at equilibrium. Using Eq. (A.25), we obtain:

$$J_{m \to s} = -A^* T^2 e^{-\frac{q\Phi_B}{k_b T}} \tag{A.26}$$

The total current density is therefore:

$$\begin{aligned} J &= J_{s \to m} + J_{m \to s} = A^* T^2 e^{-\frac{q\Phi B}{k_b T}} \left[e^{\frac{qV}{k_b T}} - 1\right] \\ &= J_{ST} \left[e^{\frac{qV}{k_b T}} - 1\right] \end{aligned} \tag{A.27}$$

This expression shows that the thermionic emission current resembles the diode equation obtained in Eq. 9.52. The difference lies in the saturation current density which is now given by:

$$J_{ST} = A^* T^2 e^{-\frac{q\Phi_B}{k_b T}} \tag{A.28}$$

References

Bethe HA (1942) Theory of the boundary layer of crystal rectifiers. MIT Radiation Laboratory Report 43–12
Sze SM (1981) Physics of semiconductor devices. Wiley, New York

A.10 Physical Properties and Safety Information of Metalorganics

Table A.2 and Table A.3 summarize some of the basic thermodynamic properties of metalorganic sources commonly used in MOCVD, including their chemical formula and abbreviation, boiling point, melting point, and the expression of their vapor pressure as a function of temperature.

Additional information on their other important physical properties is also provided for a number of important metalorganic sources, including diethylzinc (Table A.4), trimethylindium (Table A.5), triethylindium(Table A.6), trimethylgallium (Table A.7), and triethylgallium (Table A.8).

In the rest of this Appendix, general information about the safety of metalorganic compounds will be given. This will be helpful in developing safety and health procedures during their handling.

Chemical Reactivity
Metalorganics catch fire if exposed to air, react violently with water and any compound containing active hydrogen, and may react vigorously with compounds containing oxygen or organic halide.

Stability
Metalorganics are stable when stored under a dry, inert atmosphere and away from heat.

Fire Hazard
Metalorganics are spontaneously flammable in air, and the products of combustion may be toxic. Metalorganics are pyrophoric by the paper char test used to gauge pyrophoricity for transportation classification purposes (Mudry 1975).

Firefighting Technique
Protect against fire by strict adherence to safe operating procedures and proper equipment design. In case of fire, immediate action should be taken to confine it. All lines and equipment which could contribute to the fire should be shut off. As in any fire, prevent human exposure to fire, smoke, or products of combustion. Evacuate nonessential personnel from the fire area.

The most effective fire extinguishing agent is dry chemical powder pressurized with nitrogen. Sand, vermiculite, or carbon dioxide may be used. *Caution:*

Table A.2 Physical properties of some organometallics used in MOCVD (Ludowise 1985 and http://electronicmaterials.rohmhaas.com)

Compound	Formula	Abbreviation	Melting point (°C)	Boiling point (°C)	$\text{Log}_{10}\ P$(mmHg) (T in K)	Temperature range (°C)
Group II sources						
Dimethylberyllium	$(CH_3)_2Be$	DMBe				
Diethylberyllium	$(C_2H_5)_2Be$	DEBe	12	194	7.59–2200/T	
Bis-cyclopentadienyl magnesium	$(C_5H_5)_2Mg$	Cp_2Mg	176		25.14–2.18 ln T–4198/T	
Group IIB sources						
Dimethylzinc	$(CH_3)_2Zn$	DMZn	−42	46	7.802–1560/T	
Diethylzinc	$(C_2H_5)_2Zn$	DEZn	−28	118	8.280–2190/T	
Dimethylcadmium	$(CH_3)_2Cd$	DMCd	−4.5	105.5	7.764–1850/T	
Group III sources						
Trimethylaluminum	$(CH_3)_3Al$	TMAl	15.4	126	7.3147–1534.1/(T-53)	17–100
Triethylaluminum	$(C_2H_5)_3Al$	TEAl	−58	194	10.784–3625/T	110–140
Trimethylgallium	$(CH_3)_3Ga$	TMGa	−15.8	55.7	8.07–1703/T	
Triethylgallium	$(C_2H_5)_3Ga$	TEGa	−823	143	8.224–2222/T	50–80
Ethyldimethylindium	$(CH_3)_2(C_2H_5)In$	EDMIn	5.5			10–38
Trimethylindium	$(CH_3)_3In$	TMIn	88.4	133.8	10.520–3014/T	
Triethylindium	$(C_2H_5)_3In$	TEIn	−32	184	1.2	44
					3	53
					12	83

Table A.3 Physical properties of some organometallics used in MOCVD (Ludowise 1985 and http://electronicmaterials.rohmhaas.com)

Compound	Formula	Abbreviation	Melting point (°C)	Boiling point (°C)	Log_{10} P(mmHg) (T in K)	Temperature range (°C)
Group IV sources						
Tetramethylgermanium	$(CH_3)_4Ge$	TMGe	−88	43.6	139	0
Tetramethyltin	$(CH_3)_4Sn$	TMSn	−53	78	7.495–1620/T	
Tetraethyltin	$(C_2H_5)_4Ge$	TESn	−112	181		
Group V sources						
Diethylarsine hydride	$(C_2H_5)_2AsH$	DEAs			7.339–1680/T	
Tertiarybutylarsine	$(C_4H_9)AsH_2$	TBAs			7.5–1562.3/T	
Tertiarybutylphosphine	$(C_4H_9)PH_2$	TBP			7.586–1539/T	
Trimethylphosphorus	$(CH_3)_3P$	IMP	−85	37.8	7.7329–1512/T	
Triethylphosphorus	$(C_2H_5)_3P$	TEP	−88	127	7.86–2000/T	18–78.2
Trimethylarsenic	$(CH_3)_3As$	TMAs	−87.3	50–52	7.7119–1563/T	
Triethylarsenic	$(C_2H_5)_3As$	TEAs	−91	140	15.5	37
Trimethylantimony	$(CH_3)_3Sb$	TMSb	−86.7	80.6	7.7280–1709/T	
Triethylantimony	$(C_2H_5)_3Sb$	TESb	−98	116	17	75
Group VI sources						
Diethylselenide	$(C_2H_5)_2Se$	DESe	–	108		
Dimethyltellurium	$(CH_3)_2Te$	DMTe	10	82	7.97–1865/T	
Diethyltellurium	$(C_2H_6)_2Te$	DETe	–	137–138	7.99–2093/T	

Table A.4 Chemical properties of diethylzinc (Razeghi 1989)

Acronym	DEZn
Formula	$(C_2H_5)_2Zn$
Formula weight	123.49
Metallic purity	99.9999 wt% (min) zinc
Appearance	Clear, colorless liquid
Density	1.198 g·ml^{-1} at 30 °C
Melting point	−30 °C
Vapor pressure	3.6 mmHg at 0 °C 16 mmHg at 25 °C 760 mmHg at 117.6 °C
Behavior toward organic solvents	Completely miscible, without reaction, with aromatic and saturated aliphatic and alicyclic hydrocarbons. Forms relatively unstable complexes with simple ethers, thioethers, phosphines, and arsines but more stable complexes with tertiary amines and cyclic ethers
Stability in air	Ignites on exposure (pyrophoric)
Stability in water	Reacts violently, evolving gaseous hydrocarbons, carbon dioxide, and water
Storage stability	Stable indefinitely at ambient temperatures when stored in an inert atmosphere

Table A.5 Chemical properties of trimethylindium (Razeghi 1989)

Acronym	TMIn
Formula	$(CH_3)_3In$
Formula weight	159.85
Metallic purity	99.999 wt% (min) indium
Appearance	White, crystalline solid
Density	1.586 g·ml^{-1} at 19 °C
Melting point	89 °C
Boiling point	135.8 °C at 760 mmHg 67 °C at 12 mmHg
Vapor pressure	15 mmHg at 41.7 °C
Stability in air	Pyrophoric, ignites spontaneously in air
Solubility	Completely miscible with most common solvents
Storage stability	Stable indefinitely when stored in an inert atmosphere

Re-ignition may occur. *Do not use water, foam, carbon tetrachloride, or chlorobromomethane* extinguishing agents, as these materials react violently and/or liberate toxic fumes on contact with metalorganics.

When there is a potential for exposure to smoke, fumes, or products of combustion, firefighters should wear full-face positive-pressure self-contained breathing apparatus or a positive-pressure supplied-air respirator with escape pack and impervious clothing including gloves, hoods, aluminized suits, and rubber boots.

Table A.6 Chemical properties of triethylindium (Razeghi 1989)

Acronym	TEIn
Formula	$(C_2H_5)_3In$
Formula weight	202.01
Metallic purity	99.9999 wt% (min) indium
Appearance	Clear, colorless liquid
Density	1.260 g·ml^{-1} at 20 °C
Melting point	−32 °C
Vapor pressure	1.18 mmHg at 40 °C 4.05 mmHg at 60 °C 12.0 mmHg at 80 °C
Behavior toward organic solvents	Completely miscible, without reaction, with aromatic and saturated aliphatic and alicyclic hydrocarbons. Forms complexes with ethers, thioethers, tertiary amines, phosphines, arsines, and other Lewis bases
Stability in air	Ignites on exposure (pyrophoric)
Stability in water	Partially hydrolyzed, loses one ethyl group with cold water
Storage stability	Stable indefinitely at ambient temperatures when stored in an inert atmosphere

Table A.7 Chemical properties of trimethylgallium (Razeghi 1989)

Acronym	TMGa
Formula	$(CH_3)_3In$
Formula weight	114.82
Metallic purity	99.9999 wt% (min) gallium
Appearance	Clear, colorless liquid
Density	1.151 g·ml^{-1} at 15 °C
Melting point	−15.8 °C
Vapor pressure	64.5 mmHg at 0 °C 226.5 mmHg at 25 °C 760 mmHg at 55.8 °C
Behavior toward organic solvents	Completely miscible, without reaction, with aromatic and saturated aliphatic and alicyclic hydrocarbons. Forms complexes with ethers, thioethers, tertiary amines, tertiary phosphines, tertiary arsines, and other Lewis bases
Stability in air	Ignites on exposure (pyrophoric)
Stability in water	Reacts violently, forming methane and Me_2GaOH or $[(Me_2Ga)_2O]_x$
Storage stability	Stable indefinitely at ambient temperatures when stored in an inert atmosphere

Human Health

Metalorganics cause severe burns and should not get in the eyes and on the skin or clothing.

Ingestion and inhalation. Because of the highly reactive nature of metalorganics with air and moisture, ingestion is unlikely.

Skin and eye contact. Metalorganics react immediately with moisture on the skin or in the eye to produce severe thermal and chemical burns.

Table A.8 Chemical properties of triethylgallium (Razeghi 1989)

Acronym	TEGa
Formula	$(C_2H_5)_3Ga$
Formula weight	156.91
Metallic purity	99.9999 wt% (min) gallium
Appearance	Clear, colorless liquid
Density	1.0586 g·ml^{-1} at 20 °C
Melting point	−82.3 °C
Vapor pressure	16 mmHg at 43 °C 62 mmHg at 72 °C 760 mmHg at 143 °C
Behavior toward organic solvents	Completely miscible, without reaction, with aromatic and saturated aliphatic and alicyclic hydrocarbons. Forms complexes with ethers, thioethers, tertiary amines, tertiary phosphines, tertiary arsines, and other Lewis bases
Stability in air	Ignites on exposure (pyrophoric)
Stability in water	Reacts vigorously, forming ethane and Et_2GaOH or $[(Et_2Ga)_2O]_x$
Storage stability	Stable indefinitely at room temperatures in an inert atmosphere

First Aid

If contact with metalorganics occurs, immediately initiate the recommended procedures below. Simultaneously contact a poison center, a physician, or the nearest hospital. Inform the person contacted of the type and extent of exposure, describe the victim's symptoms, and follow the advice given.

Ingestion. Should metalorganics be swallowed, immediately give several glasses of water but do not induce vomiting. If vomiting does occur, give fluids again. Have a physician determine if condition of patient will permit induction of vomiting or evacuation of the stomach. Do not give anything by the mouth to an unconscious or convulsing person.

Skin contact. Under a safety shower, immediately flush all affected areas with large amounts of running water for at least 15 min. Remove contaminated clothing and shoes. Do not attempt to neutralize with chemical agents. Get medical attention immediately. Wash clothing before reuse.

Eye contact. Immediately flush the eyes with large quantities of running water for a minimum of 15 min. Hold the eyelids apart during the flushing to ensure rinsing of the entire surface of the eyes and lids with water. Do not attempt to neutralize with chemical agents. Obtain medical attention as soon as possible. Oils or ointments should not be used at this time. Continue the flushing for an additional 15 min if a physician is not immediately available.

Inhalation. Exposure to combustion products of this material may cause respiratory irritation or difficulty with breathing. If inhaled, remove to fresh air. If not breathing, clear the victim's airway and start mouth-to-mouth artificial respiration which may be supplemented by the use of a bag-mask respirator or manually triggered oxygen supply capable of delivering 1 liter per second or more. If the

victim is breathing, oxygen may be delivered from a demand-type or continuous-flow inhaler, preferably with a physician's advice. Get medical attention immediately.

Industrial Hygiene

Ingestion. As a matter of good industrial hygiene practice, food should be kept in a separate area away from the storage/use location. Smoking should be avoided in storage/use locations. Before eating, hands and face should be washed.

Skin contact. Skin contact must be prevented through the use of fire-retardant protective clothing during sampling or when disconnecting lines or opening connections. Recommended protection includes a full-face shield, impervious gloves, aluminized polyamide coat, hood, and rubber boots. Safety showers – with quick-opening valves which that stay open – should be readily available in all areas where the material is handled or stored. Water should be supplied through insulated and heat-traced lines to prevent freeze-ups in cold weather.

Eye contact. Eye contact with liquid or aerosol must be prevented through the use of a full-face shield selected with regard for use-condition exposure potential. Eyewash fountains, or other means of washing the eyes with a gentle flow of tap water, should be readily available in all areas where this material is handled or stored. Water should be supplied through insulated and heat-traced lines to prevent freeze-ups in cold weather.

Inhalation. Metalorganics should be used in a tightly closed system. Use in an open (e.g., outdoor) or well-ventilated area to minimize exposure to the products of combustion if a leak should occur. In the event of a leak, inhalation of fumes or reaction products must be prevented through the use of an approved organic vapor respirator with dust, mist, and fume filter. Where exposure potential necessitates a higher level of protection, use a positive-pressure, supplied-air respirator.

Spill Handling

Make sure all personnel involved in spill handling follow proper firefighting techniques and good industrial hygiene practices. Any person entering an area with either a significant spill or an unknown concentration of fumes or combustion products should wear a positive-pressure, supplied-air respirator with escape pack. Block off the source of spill, and extinguish fire with extinguishing agent. Re-ignition may occur. If the fire cannot be controlled with the extinguishing agent, keep a safe distance, protect adjacent property, and allow product to burn until consumed.

Corrosivity to Materials of Construction

This material is not corrosive to steel, aluminum, brass, nickel, or other common metals when blanketed with a dry inert gas. Some plastics and elastomers may be attacked.

Storage Requirements

Containers should be stored in a cool, dry, well-ventilated area. Store away from flammable materials and sources of heat and flame. Exercise due caution to prevent damage to or leakage from the container.

References

Ludowise M (1985) Metalorganic chemical vapor deposition of III-V semiconductors. J Appl Phys 58:R31–R55

Mudry WL, Burleson DC, Malpass DB, Watson SC (1975) J Fire Flammability 6:478.

Razeghi M (1989) The MOCVD challenge Volume 1: a survey of GaInAsP-InP for photonic and electronic applications. Adam Hilger, Bristol

Sze SM (1981) Physics of semiconductor devices. Wiley, New York

Index

M. Razeghi, *Fundamentals of Solid State Engineering*,
https://doi.org/10.1007/978-3-319-75708-7